DOCUMENT RESUME

ED 226 414 EA 015 336

AUTHOR Lazarus, Mitchell
TITLE Evaluating Educational Programs.
INSTITUTION American Association of School Administrators, Arlington, Va.
PUB DATE 82
NOTE 71p.
AVAILABLE FROM Publications, American Association of School Administrators, 1801 North Moore Street, Arlington, VA 22209 (Stock No. 021-00852; $10.95; quantity discounts).
PUB TYPE Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Achievement Tests; Educational Environment; Educational Improvement; Educational Objectives; *Educational Testing; Elementary Secondary Education; *Evaluation Methods; Evaluation Utilization; Guidelines; Institutional Evaluation; Minimum Competency Testing; *Program Evaluation; Statistical Analysis; Student Evaluation; Testing Problems; Test Interpretation

ABSTRACT

In eight chapters, this document offers guidelines for school administrators on evaluating schools and school programs, based on a poll of administrators and consultants. The introduction discusses the purpose of evaluation, especially school improvement, and presents a list of evaluation standards. Chapter 1 outlines steps in evaluation planning, including defining the objectives and audience, choosing techniques, and reporting results. Statistical evaluation is introduced in chapter 2, which covers statistical significance, extraneous factors, and pre- and posttests. Chapter 3 explores choosing tests appropriate for the evaluation purposes, taking note of test reliability, errors, validity, item analysis, and bias. Test administration and the elimination of environmental influences, such as the effects of physical surroundings, anxiety, and test-taking skills, are addressed in chapter 4. Chapter 5 discusses interpreting test scores and explains norm- and criterion-referenced tests, averages, percentiles, and inspection of test results. Arguments concerning minimum competency testing are summarized in chapter 6, while a number of emerging evaluation methods that eschew tests, such as qualitative assessments, videotaping, and interviews, are described in chapter 7. Chapter 8 relates evaluation to current technological trends and school improvement and accreditation. Sample evaluation policies and a glossary of 68 evaluation-related terms are appended. (RW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

PROGRAMS

ED226414

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

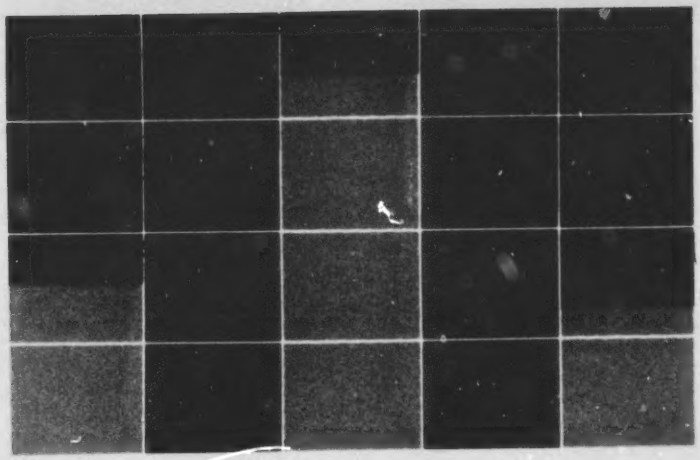
"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

W. Henry

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

EA 015 336

EVALUATING EDUCATIONAL PROGRAMS



By Mitchell Lazarus

Publications from the American Association of School Administrators

Evaluation Series
Evaluating Educational Personnel
Evaluating Educational Programs

Superintendent Career Development Series
Books in this series include:
Guidelines for the Preparation of School Administrators
Selecting a Superintendent
The Superintendent's Contract
Compensating the Superintendent (Summary Report)
Compensating the Superintendent (Full Report)
The Administrative Leadership Team
Evaluating the Superintendent

Administrative Team Development Series
Books in this series include:
Selecting the Administrative Team
The Board's Role in Selecting the Administrative Team
Classifying and Compensating the Administrative Team
Classifying and Compensating the Administrative Team (Full Report)

Critical Issues Reports
Books in this series include:
School Budgeting: Problems and Solutions
Declining Enrollment/Closing Schools: Problems and Solutions
Collective Bargaining: Problems and Solutions
Student Discipline: Problems and Solutions
Keeping Students in School: Problems and Solutions
How Schools Can Save Dollars: Problems and Solutions
Staff Dismissal: Problems and Solutions
School Energy Crisis: Problems and Solutions
The Competency Movement: Problems and Solutions

Cassette Tapes
A four-cassette instructional package:
How to Evaluate the Performance of Educational Professionals
School Board Member Self-Evaluation
Evaluating the Superintendent
Evaluating Principals and Central Office Administrators

These materials are available from the American Association of School Administrators, 1801 North Moore Street, Arlington, Virginia 22209. Price lists available. Discounts on multiple copy orders.

Stock Number: 021-00852
Library of Congress Card Catalogue Number: 81-70822
Copyright, 1982 The American Association of School Administrators, Arlington, Virginia

Design by Kim K. Hasten Graphic Design

Points of view or opinions in this book are those of the author and contributors and do not necessarily represent the official position or policies of the American Association of School Administrators.

Table of Contents

Foreword 5

Introduction 7

Chapter 1
Evaluation Planning:
Step-by-Step Effectiveness 13

Chapter 2
Statistical Evaluation:
Techniques Using Test Scores 21

Chapter 3
Choosing a Test:
The Right One for the Purpose 29

Chapter 4
Using a Test:
Eliminating Environmental Influences 39

Chapter 5
Interpreting Test Scores:
Looking Beyond the Numbers 45

Chapter 6
Minimum Competency Testing:
A Special Method of Evaluation 53

Chapter 7
Emerging Evaluation Methods:
Emphasizing Students over Numbers 57

Chapter 8
Evaluation to Improvement:
A Never-Ending Process 65

Sample Policies 73

Glossary 75

Acknowledgments 79

Foreword

Educational quality. Our communities expect it. Our students deserve no less than the highest quality, most effective education we can offer.

But how can we assure that the educational programs of our nation's schools are effective? How can we be assured that those programs respond to the real needs of our communities and of society? One way to measure effectiveness is through a well developed system for evaluation.

Evaluation of educational programs is an inexact science. As you review this publication, you will find that most school systems rely on testing to determine a program's effectiveness. Testing has both advantages and limitations. And it is important for us to be keenly aware of the idiosyncracies of testing.

Many school systems combine testing information with other evaluation methods. School improvement councils, self studies, visits by external evaluation teams, examinations of whether existing programs are meeting real needs, all contribute to the evolving science, or art, of effective evaluation.

The American Association of School Administrators (AASA) is committed to the improvement of learning. That goal must become our national imperative. Educators must take the lead in assuring the most effective schools possible.

In preparing this publication, AASA polled a representative sample of school administrators across the nation. Those educational leaders shared their ideas, insights and techniques which are included in this important book. *Evaluating Educational Programs* is the second in a two part series devoted to the improvement of our schools. The first book in this series, *Evaluating Educational Personnel*, provides important information about evaluating the people who make education happen.

Both of these publications take a vital, but sometimes missing step. They describe how the results of an evaluation program can and should be applied to assure educational improvement. Making our schools better tomorrow than they are today . . . should be the ultimate goal for any effective system of evaluation.

Paul B. Salmon, Ed.D.
Executive Director, AASA

Introduction

School systems across the nation face mounting demands on their steadily declining resources. From the smallest rural districts to the districts of the largest cities, administrators continually must make hard decisions on the allocation of increasingly scarce personnel and funds. When the issues calling for decision include instructional programs—and they often do—program evaluation is one important step in the decision-making process.

"Unfortunately, there has not been sufficient use of evaluation results when the decisions are made to reduce programs. A decline in student enrollment is normally used as the major rationale along with adherence to union contracts," says Donald P. Mys, consultant to the Dearborn, Michigan, public schools.

The role of program evaluation is not, however, limited to situations demanding specific or immediate action. Program evaluation is equally important as a standing component in the everyday operation of a school system.

Improving Education: The Major Purpose of Evaluation

"The major purpose of any program evaluation is the improvement of the instructional process," Mys adds.

"All evaluation should relate to program improvement," echoes James Nighswander, director of instruction for the Springfield, Illinois, public schools.

Instructional programs that were satisfactory a few years ago may not work as well today. Students change; teachers change; local needs change.

Old programs may need revision or replacement. Programs developed more recently may be more suited to the job and lead to educational improvement. The routine evaluation of ongoing practices can detect changing needs before they accumulate into crises.

"You evaluate in order to plan for change, which hopefully, is an improvement," explains L.E. Buzzomo, superintendent of the Lower Moreland School District in Huntingdon Valley, Pennsylvania.

This publication, *Evaluating Educational Programs*, explores two parts of the program evaluation issue: what the schools are doing and what they should be doing. A companion report, *Evaluating Educational Personnel*, looks at how the issue of personnel evaluation affects instructional programs. Often, these two areas overlap.

Working Together: Curriculum and Evaluation

"We have a close working relationship between the curriculum and evaluation departments. The curriculum department uses evaluation results for instructional change each summer," says Don R. Robinson, assistant superintendent for research, planning and eval-

uation for the Duval County, Florida, public schools.

Sometimes, especially in smaller districts, those who plan and administer evaluations and those who make decisions using the results are the same. But more often, action resulting from an evaluation is taken by administrators who may have limited expertise in evaluation.

Sometimes a language barrier develops when decision makers are not completely familiar with evaluation terminology and its nuances.

To provide some help in the interpretation of evaluation results, this report gives some of the more important technical background in plain language. In addition, a glossary on page 75 will help explain the technical terms that are used.

Based on the information gleaned from a survey of members of the American Association of School Administrators (AASA), it is clear that districts evaluate for a variety of purposes and use a broad range of techniques. Some 213 school districts replied to a five-page questionnaire about their evaluation efforts. Many of those districts also provided reports and other materials that illustrated their own practices.

Responding districts vary widely in size. The average number of students they serve is 10,000 with the range extending from a few hundred students to beyond 100,000. The typical number of students in the replying districts is a few thousand. The number of schools in each district ranges from one to more than 100 with the average number of schools in the replying districts about 15.

According to the survey responses, some districts pay little attention to evaluation. Others evaluate only when required to meet some special need. Still other districts view program evaluation as an important ongoing activity given considerable support.

District Size: A Significant Factor

Perhaps the most important single factor in the organization and extent of evaluation efforts in a school district is the size of the district. An urban district or a large county typically has an entire department devoted to evaluation, with much of the effort given over to program evaluation. The staff usually consists, at least in part, of people with post-graduate training in educational testing and measurement. Some staff members often are free to spend considerable time in the schools, working directly with teachers and children. Others, using computers dedicated to the purpose, conduct elaborate analyses and breakdowns of trends. The reports from such an evaluation team often display considerable professional and mathematical sophistication.

At the other extreme, a small, rural district consisting of only a few schools cannot usually afford even one part-time employee to work at evaluation. There, the evaluation tasks must often be shared among teachers and

Evaluation Goals

The Evaluation Research Society identifies the following five goals for program evaluation, in addition to the goal of assessing a program's impact:

1. **Front end analysis**—activities taking place prior to the initiation of a program to determine necessary levels of support or even whether to initiate the program at all;
2. **Evaluability Assessment**—an analysis of the feasibility of conducting future evaluation activities such as an impact study;
3. **Formative evaluation**—appraising the process of a program, particularly during a period of program development in order to improve its functioning;
4. **Program monitoring**—the periodic examination of programs that might focus on issues of program compliance or the collection of basic descriptive data that might assist in program management;
5. **Evaluation of an evaluation**—this may take a variety of forms ranging from a professional critique of a report to reanalysis of original data, to the collection of new data.

8

principals who already have more than enough to do. Such districts often must manage without the fancy analyses that require computer equipment and specially trained people. But with fewer children to examine, it is often possible to evaluate the children more individually. In many ways, this leads to a more satisfactory evaluation than one based on a mountain of statistics.

Indeed, good evaluation seems to follow not so much from extensive training and lavish budgets as from a genuine sensitivity to students, their accomplishments and their needs. That, fortunately, is a characteristic shared by many districts large and small, rich and poor.

Setting Standards: A Guide for Professionals

Recent years have seen much promulgation of evaluation "standards." One set was developed by the Joint Committee on Standards for Educational Evaluation. The committee was chaired by Daniel L. Stufflebeam and included members from 12 professional organizations of school administrators, board members, researchers and teachers, including AASA. Designed as a guide for those who commission, conduct and employ evaluations of educational programs, the publication, *Standards for Evaluations of Educational Programs, Projects and Materials*, discusses 30 separate standards. (See illustration, "Standards for Evaluation on page 10.)

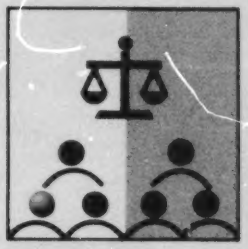
Professional journals also have urged the establishment of written evaluation "policies." But all such formalities are merely a means to an end. They encourage the carrying out of evaluations with respect, intelligence and sensitivity. Published policies and standards can spell out certain minimums; and at a minimum, such standards are certainly necessary. But once more, the success of an evaluation effort depends most of all on the people involved. People who care enough to watch and listen attentively to students will, as a rule, conduct good evaluations regardless of the policies and standards under which they operate.

This report can help to provide some raw material for such people. It discusses some of the factors involved in planning an evaluation, in carrying it out and in interpreting and making use of the results. Much of the emphasis here is on testing—and much of it concerns the drawbacks of testing as a primary means of evaluation.

Standardized Testing: A Dominant Practice Nationwide

Standardized testing by far dominates the practice of program evaluation nationwide. Much of the testing is carried out uncritically, with little thought to its limitations. There is a growing body of professional opinion to the effect that standardized testing, while usually the fastest and cheapest way to evaluate, is sometimes not the best way. Accordingly, some of the pitfalls in testing and how to avoid them are discussed. This report suggests some practical alternatives, methods that have been proven to work.

Among the criteria for good evaluation is good communication—between learners and evaluators, first of all, and also between evaluators and teachers, principals, curriculum personnel, school boards and parents. An evaluation can be of no help unless its results find use in practice. For that, communication is essential. The large districts have "channels"—boards and panels and formal review mechanisms—to ensure that evaluators' views are heard. Small districts may have to accomplish the same thing amidst other business at staff meetings, over the coffeepot in the teachers' lounge or through newsletters to parents. But again, mechanism is less important than results. Regardless of size and resources, a district whose personnel genuinely care about educational quality will always find ways to improve it.



Standards for Evaluation

To ensure that an evaluation will serve the practical information needs of given audiences:

- Audiences affected by the evaluation should be identified so their needs can be addressed.
- Those conducting the evaluation should be both trustworthy and competent to perform it so their findings achieve maximum credibility and acceptance.
- Information collected should be of such scope and selected in such ways as to address pertinent questions about the object of the evaluation and be responsive to the needs of specified audiences.
- The perspectives, procedures and rationale used to interpret the findings should be carefully described so that the bases for value judgments are clear.
- The evaluation report should describe the object being evaluated and its context, and the purposes, procedures and findings of the evaluation, so that the audiences will readily understand what was done, why it was done, what information was obtained, what conclusions were drawn and what recommendations were made.
- Evaluation findings should be disseminated to audiences so they can assess and use the findings.
- Release of reports should be timely so audiences can best use the reported information.
- Evaluations should be planned and conducted in ways that encourage follow-through by the audiences.

To ensure that an evaluation will be realistic, prudent, diplomatic and frugal:

- The evaluation procedures should be practical so disruption is kept to a minimum and needed information can be obtained.
- The evaluation should be planned and conducted with the anticipation of the different positions of various interest groups so their cooperation may be obtained and so possible attempts by any of these groups to curtail evaluation operations or to bias or misapply the results can be averted.
- The evaluation should produce information of sufficient value to justify the resources expended.

To ensure that an evaluation will be conducted legally, ethically and with due regard for the welfare of those involved in the evaluation, as well as those affected by its results:

- Obligations of the formal parties to an evaluation (what is to be done, how, by whom, when) should be agreed to in writing so these parties are obligated to adhere to the agreement or to renegotiate it.
- Conflict of interest, frequently unavoidable, should be dealt with openly and honestly so it does not compromise the evaluation process and results.
- Oral and written evaluation reports should be open, direct and honest in their disclosure of pertinent findings, including the limitations of the evaluation.
- The formal parties to an evaluation should

respect and assure the public's right to know, within the limits of other related principles and statutes, such as those dealing with public safety and the right to privacy.

- Evaluations should be designed and conducted so the rights and welfare of the human subjects are respected and protected.
- Evaluators should respect human dignity and worth in their interactions with other persons associated with an evaluation.
- The evaluation should be complete and fair in its presentation of strengths and weaknesses of the object under investigation so strengths can be built upon and problem areas addressed.
- The evaluator's allocation and expenditure of resources should reflect sound accountability procedures and otherwise be prudent and ethically responsible.

To ensure that an evaluation will reveal and convey technically adequate information about the features of the object being studied that determine its worth or merit:

- The object of the evaluation (program, project, material) should be sufficiently examined so the form(s) of the object being considered in the evaluation can be clearly identified.
- The context in which the program, project or material exists should be examined in enough detail so its likely influences on the object can be identified.
- The purposes and procedures of the evaluation should be monitored and described in enough detail so they can be identified and assessed.
- The sources of information should be described in enough detail so the adequacy of the information can be assessed.
- The information-gathering instruments and procedures should be chosen or developed and then implemented in ways that will assure that the interpretation arrived at is valid for the given use.
- The information-gathering instruments and procedures should be chosen or developed and then implemented in ways that will assure the information obtained is sufficiently reliable for the intended use.
- The data collected, processed and reported in an evaluation should be reviewed and corrected so the results of the evaluation will not be flawed.
- Quantitative information in an evaluation should be appropriately and systematically analyzed to ensure supportable interpretations.
- The conclusions reached in an evaluation should be explicitly justified so an audience can assess them.
- The evaluation procedures should provide safeguards to protect the evaluation findings and reports against distortion by the personal feelings and biases of any party to the evaluation.

Source: Excerpted from *Standards for Evaluations of Educational Programs, Projects, and Materials*, developed by the Joint Committee on Standards for Educational Evaluation, Published by McGraw-Hill Book Company. Used by permission.

Chapter 1

Evaluation Planning: Step-by-Step Effectiveness

As with almost every other endeavor, the key to effective program evaluation lies in planning. "You must set objectives if you evaluate," reports Bill Kennedy, superintendent of the Cedar Hill Independent School District in Cedar Hill, Texas.

One excellent planning outline was developed by district personnel in Dearborn, Michigan, using a model they credit to the California State Department of Education:

1. Determine evaluation purpose and requirements.
2. Develop an evaluation plan.
3. Determine evaluation design and do the sampling.
4. Select or develop assessment instruments.
5. Collect the data.
6. Analyze evaluation data.
7. Report evaluation findings.
8. Apply evaluation findings.

Even the smallest evaluation must go through all of these steps. When an evaluation becomes large enough to require outside help, the planning becomes correspondingly more important.

The Joint Committee on Standards for Educational Evaluation offers the following checklist for evaluation planning.

- Developing evaluation policies;
- Deciding whether to evaluate;
- Defining the evaluation problem;
- Writing a contract to cover and control an evaluation;
- Designing an evaluation;
- Budgeting for an evaluation;
- Staffing an evaluation;
- Training evaluators;
- Administering and monitoring evaluation operations;
- Producing and communicating evaluation reports;
- Evaluating an evaluation;
- Deciding what to do with evaluation results.

Deciding on a Purpose: A Guide to the Process

An evaluation's ultimate purposes ought to guide every other stage of the work: the evaluation plan and design, choice of instruments, analysis and form of report. Decisions on purposes—on precisely what the evaluation seeks to find out—can be the most important decisions in the entire course of the evaluation. They set the course for the other decisions that follow.

Sometimes, however, the purposes go largely unspecified. Asked one evaluator, hunched over a stack of computer printouts, "What is the question to which all this is the answer?"

Thinking through purposes is hard work.

The reasoning inevitably leads to very difficult questions. Do we care if students are learning this? Why do we care? Should we be teaching this at all? What should we be teaching? What is most important for students to learn? What is the right intellectual equipment for adult life? What kinds of lives might people lead? What is worthwhile?

With disarming ease, a simple evaluation slides into profound philosophical issues. The problem of choosing a reading test reflects, after a little thought, a choice of educational goals. Those in turn come to rest on the kinds of people students eventually will be, a question that challenges basic beliefs.

Yet these are among the questions that educators must face. These questions help make teaching a worthy and honorable profession. And it is these questions that program evaluations must help to answer, as best they can. "It does seem that it is difficult to let go of the past and search for that unknown, untested program that could more adequately provide for the future," says Ernest Seifried, superintendent of schools, Saranac Lake Central District in New York.

Skipping a Step: Tempting but not Practical

It can be very tempting to slip over the question of purposes: "Let's not worry about that now. Let's just get the data and see what it says later." There are two serious problems with this position. First, the data will depend to a large extent on the tests used, and a sensible choice of test can be made only with particular purposes in mind. Second, the analysis—"what the data says"—likewise depends on the purposes for which the data are analyzed. Ordinarily an analysis will answer only those questions that are asked. Sometimes there emerge results so striking they need not be sought, but this is the exception.

Program evaluation has a host of possible purposes. Here is a partial list of broad categories. Each category would need to be more specific in order to be useful in any particular case.

- Planning curriculum overall (many subjects; many grades);
- Choosing broad instructional goals;
- Specifying objectives for particular programs;
- Detecting the need for program revision or replacement;
- Designing new programs or revisions;
- Allocating resources (time, people, money, space);
- Minimizing the damage from cutbacks, school closings, reductions in force;
- Responding to federal mandates;
- Meeting state and local requirements;
- Responding to political pressure;
- Maintaining sound public relations.

Choosing Goals: The Fewer the Better

No one evaluation can accomplish all of these goals, or even very many of them. As a rule, an evaluation should tell the most about the least. In fact, the fewer the questions asked, the more accurately and thoroughly they can be answered. Thus it is usually desirable to hold the number of issues to a minimum.

An evaluator will often be under pressure to expand the scope of an evaluation. Considering the expense and disruption involved, there is a natural tendency to make the most of it. "As long as we have to do this for the state anyway, can't we run a check on that new math program at the same time?" Sometimes a little extra work will greatly increase the evaluation's payoff, by providing extra information. There is a risk in trying to do too much however. The operation can become hopelessly unwieldy, and the data in the end may not be fully satisfactory for any of the intended purposes.

Two or three small evaluations can be simpler, more informative, and possibly even less expensive than one big one. In Oshkosh, Wisconsin, evaluations vary in size and scope. "Program evaluations are conducted by both district staff and outsiders. We conduct an internal needs assessment and evaluation first. Then the external team validates the internal study and makes recommendations. We expect ideal situations and recommendations to be presented by the external team to enable the district to grow," explains Joseph Pellegrin, superintendent in Oshkosh.

Writing it Down: One Way To Promote Consistency

When a district is serious about setting specific purposes for its evaluations, it makes sense to establish a written policy or set of guidelines. Yet fewer than a third of AASA's survey respondents do so.

In the absence of written guidelines, maintaining a consistent evaluation policy is difficult. Every new evaluation becomes to some extent an ad hoc operation. Comparisons among evaluations are difficult, especially if evaluation personnel change from year to year and the styles of evaluation change with them. Moreover, without a written policy, some instructional areas may go with no evaluation at all, until conditions have deteriorated to the point where the need for change is obvious even without one.

A written evaluation policy can usefully provide the following kinds of information:

- The district's commitment to routine, ongoing evaluation as part of its overall educational mandate;
- With whom the responsibility for program evaluation lies, and to whom those people report;
- Purposes of evaluations;
- "Audiences" for evaluations;
- Routine evaluations;
- Special evaluations;
- Preferred forms of reporting;
- General requirements for instruments to be used;
- Adherence (or not) to the *Standards for Evaluations of Educational Programs, Projects, and Materials*, promulgated by the Joint Committee on Standards for Educational Evaluation and summarized in the introduction to this publication;
- Policies on release of evaluation information to persons outside the school system, including parents and the press;
- Procedures to ensure confidentiality for individual students and teachers.

Specifying Purposes: How it's Done

The AASA survey conducted in preparation for this publication provided the following information on how districts specify purposes for evaluations. The percentages below reflect only those districts answering each particular question. The results should not be taken to apply generally to the nation.

- Only 32 percent of responding districts have a general written policy or guidelines on program evaluation. But 66 percent put the objectives and methods for each particular evaluation in writing.
- About 42 percent are subject to evaluation requirements imposed by either the state or the local board.
- About 55 percent evaluate their individual instructional programs on a routine schedule. Another 23 percent evaluate only when the need arises, and 22 percent evaluate under both circumstances.
- The intervals between routine evaluations vary widely among districts and with subject area, ranging from every few months to every seven or ten years. The average for core programs is about every two to three years. Surprisingly, one or two percent **never** evaluate even their elementary reading and math programs. Evaluations at the second-

Useful Evaluations

Useful evaluations begin with carefully reasoned evaluation plans conceived when the program is designed. For the results of an evaluation study to be most useful, one must know:

- What is expected of the evaluation;
- How and by whom the evaluation is going to be used;
- Who the primary program beneficiaries are;
- What, exactly, the program is expected to do for its clients;
- What goes into the program;
- How the program actually operates;
- How the program is accepted by clients; and
- Whether and how the clients actually benefited from the program.

Source: Assessment and Evaluation Division, Education Commission of the States

ary level average only a few months further apart than at the elementary level.

- The overall instructional program receives less attention than do the individual subject areas. Fourteen percent never evaluate overall, and 28 percent do so only "when needed." Among the 58 percent who evaluate the entire program routinely, the average interval is 3.1 years. This average may be misleading, however. A great many districts evaluate either annually or at intervals of from five to seven years.
- When asked whether "your program evaluation system is tied directly into educational improvement," fully 83 percent answered yes.
- Some 54 percent use formal program evaluations in making "program cutbacks that become necessary for economic reasons." Several respondents, however, volunteered that they had not faced the issue. "Not yet," said many.
- Seventy percent evaluate not only program effectiveness, but also for educational objectives.
- Seventy-one percent include procedures to safeguard the identities of participating students. Almost as many protect teachers

Choosing Techniques

Ronald Edmonds, professor of education in the Department of Teacher Education at Michigan State University, insists that educational programs be justified on the "best evidence available." The form of that evidence depends on two main factors: the type of subject matter and the student population. Taking both of those into account, evaluators should look for what Edmonds calls the "index of greatest confidence."

Tests are useful chiefly in subject areas that have a lot of testing experience behind them, such as reading and math. Results in some other areas, especially those that encourage more subjective goals and objectives, do not show up well on a test. In those cases, Edmonds recommends more subjective measures, including student-written essays and teacher judgments.

The case for tests becomes stronger where classes are socially mixed. Edmonds' experience shows that judgments of middle-class teachers can be unreliable in assessing students from differing socio-economic backgrounds. Some teachers, he finds, tend to be over-liberal in judging disadvantaged students, while others tend to be overcritical. Either way, the evaluation results are open to question, and a test becomes preferable. The inadvertent discrimination is less important, however, when classes are socially homogeneous.

Edmonds expects middle-class children to do better on tests, on the whole. But instead of looking at the overall scores, as most evaluators do, he concentrates on the percentage of children who attain minimum mastery. Once the child has attained that level, the amount by which his or her score exceeds the minimum is of less interest to Edmonds. He especially seeks roughly equal percentages of both socio-economic groups who meet the minimum.

The ideal instructional environment, says Edmonds, is one in which there is pervasive understanding throughout the community—"when people who trust each other talk to each other about what's going on." Often this is not the case, however. There can be hostility between those who run the schools and their clients, particularly when a middle-class teaching staff works with lower-class children. In that situation, says Edmonds, it is possible to neutralize much of the hostility by using the most objective description of pupil progress available.

as well. The most common safeguards are reporting of results by number rather than name and restricting access to the raw results to professional evaluation personnel.

Defining the Audience: Who Needs the Results

The question of purposes ties closely to the question of just who the evaluation is for. The many different potential audiences for an evaluation include:

- Classroom teachers;
- Building principals;
- District curriculum staff;
- District evaluation staff;
- Textbook committees;
- The school board;
- Local government;
- State curriculum staff;
- State evaluation staff;
- The state board of education;
- Federal agencies;
- Parents;
- The media;
- The public generally.

Each purpose for an evaluation will, as a rule, have its own collection of audiences. Because the evaluation is intended to meet specific audiences' needs, the audience must be kept in mind throughout the planning process. Just as mixing too many purposes can be unwise, so can mixing too many audiences—especially if some are laypeople and some are professionals.

Designing the Evaluation: Consider the Consumer

For example, an evaluation report for the local curriculum staff might properly contain hard criticism of the district's core program. The report might, for example, provide data showing the program could more effectively serve minority students. Typically such a report will base its findings on relatively small statistical differences drawn from large numbers of students and will include an analysis of possible errors. In all, the report will communicate clearly from one professional to another. Should it find its way into the newspapers, however, distortions can occur. It is not hard to imagine the headline: "Minority Kids Give Schools an 'F.'"

Conversely, an evaluation whose results are designed mostly for public consumption will not be of much help to professionals. The public often seems most interested in how well local averages compare with state and national averages. This kind of information can be broadly useful for determining whether further evaluations are necessary, but not for much else. The questions most important to professionals—**why** a particular program is working well or not—are of much less concern to the public, unfortunately.

The performance of local students, nonetheless, is a matter of legitimate public concern. As a result, the question of public relations plays a part in many evaluations. Sometimes evaluations are conducted with public reporting expressly in mind; sometimes it is a background concern. And some districts seem more skillful than others in effectively interpreting information for the press.

"We believe citizens have a right to know the district picture, but not individual or building results. The building staff has data on its school compared to the district as a whole. We do not release comparisons among schools," says Inabell Kirby, director of research, Decatur, Illinois, Public Schools.

Many respondents to our survey specifically volunteered that the need to communicate with the public is a factor in their evaluation planning.

Reporting the Results: Keep the Readers in Mind

Like every other stage of the evaluation process, the report must meet the needs of the audience in question. Not only must it give whatever information the audience requires, but it must do so in ways that the audience will understand.

Even if the report is intended only for professional evaluators, it is usually wise to include an "executive summary" in plain language. And conversely, if the report is meant for lay readers, there should still be a "technical appendix" that justifies the conclusions. Writing a clear, honest report that conveys the findings accurately to particular readers often

ranks with setting purposes as the most difficult part of an evaluation. It is also equally important. To its users, the evaluation can only be as good as the report.

The initial planning should provide for this final step. Advance knowledge of what information the report will have to include can be very helpful as the evaluation proceeds. By proceeding with the report in mind, and by organizing materials appropriately along the way, the task of preparing the report itself can be greatly simplified.

Making Decisions: Weighing the Variables

A decision on educational programs is nearly always a compromise. Evaluation results are only one part of the total mix of factors that administrators typically must consider. Some of these factors are listed in the illustration on this page, together with percentages of responding districts that consider them important. Another group of factors, sometimes the determining ones, are largely political in nature.

It would be wrong, even if it were possible, to give evaluation data too much weight in decision-making. For reasons discussed throughout this report, evaluation results are subject to many kinds of error. Even more important, the typical test-score evaluation is simply often not sensitive to many important aspects of an educational program. Much of what goes on in the classroom and much of what students gain from the program, does not emerge at all in the test scores. As a result, such an evaluation tells only part of the story. Adding qualitative methods can improve the situation considerably, but a good deal about the program's total effects is likely to remain unknown.

This problem takes its most acute form when programs must be cut due to limitations in enrollment, funding or both. Choosing the survivors solely by test score can result in a basic educational shift that may not serve all students well. Looking beyond the test scores then becomes especially important, and values apart from measured achievement must be given full consideration.

At the same time, however, many program evaluations have far less influence than they deserve. Two of the main obstacles are lack of confidence in the evaluation process, on the part of some administrators, and unclear reporting by evaluators. Good evaluation reports can help to solve the first problem as well as the second.

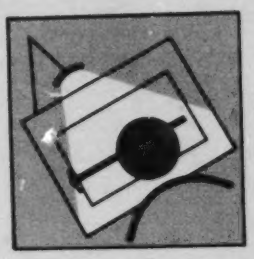
A report should be forthright about its limitations—the likely error in scores and any other sources of error that may apply. The report should present alternative interpretations of the data, together with the evaluator's opinions as to their soundness. If the evaluator has any misgivings about the quality of a test, or of its suitability for a particular purpose, these concerns should also be shared in the report. This kind of disclosure is, first of all, a professional obligation. And second, expressing one's own reservations about certain aspects of the study will tend, paradoxically, to give the reader confidence in the rest of it.

A well-organized, clearly written report is much more likely to have an impact on the final decisions than one that is hard to follow.

Factors in Decision-Making About Educational Programs

	Very Important	Important
Educational effectiveness	84%	16%
Educational objective	79%	19%
Overall expense	38%	48%
Teachers' time	35%	62%
Parents' reactions	35%	60%
Students' time	34%	61%
Need for teacher training	25%	58%
Need for special materials	22%	61%
Need for space	15%	54%
Need for transportation	10%	42%

Source: AASA survey for this publication



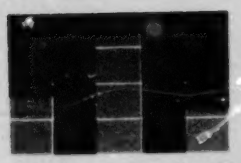
Evaluation Design

Goal Area	Evaluation Devices	Frequency
Communication skills and concepts	Administer the Comprehensive Test of Basic Skills (CTBS) at grades 2, 4, 6, 9, 11.	yearly
	Complete an objective-referenced program evaluation on a planned cycle.	once each five years
	Administer student and parent surveys in relation to program objectives--particularly "non-testable" objectives.	once each five years
	Conduct and examine time studies at the elementary level to determine opportunity and emphasis.	every other year
	List and examine secondary course offerings to determine opportunity.	every other year
Mathematics	Administer the CTBS at grades 2, 4, 6, 9, 11.	yearly
	Complete an objective-referenced program evaluation on a planned cycle.	once each five years
	Examine a 5-year graduate follow-up relative to value of courses taken.	once every five years
	Conduct elementary time studies to determine opportunity and emphasis	every other year
	List and examine secondary course offerings to determine opportunity.	every other year
	Administer parent and student surveys relative to program objectives—particularly "non-testable" objectives.	once each five years

Source: Dennis E. Rens, director of curriculum and instruction, District 13 Schools, Columbia Heights, Minnesota

Chapter 2

Statistical Evaluation: Techniques Using Test Scores



In serving a wide range of purposes, program evaluations take a wide variety of forms that range from simple to very complex. Most are simple. For example, suppose a teacher has just completed a unit on adding two-digit numbers with regrouping. To find out if the students understood the process, the teacher might run off a simple, ten-item test. A glance at the results will show fairly well if the topic needs further work, or if the class can safely go ahead and build on the completed unit. Thousands of these evaluations occur every day. Though fast and informal, they meet the requirements of a program evaluation as fully as any major study.

Other, more thorough evaluations include the National Assessment of Educational Progress (NAEP), a massive effort funded by the National Institute of Education and conducted by the Education Commission of the States (ECS). NAEP runs a program evaluation for the entire United States, seeking to detect any changes in the outcomes of education nationwide. NAEP carefully chooses 70,000 to 100,000 students in three age groups each year and tests them in ten subject areas. The outcome is a series of voluminous (and fascinating) reports that track student performance, skill by skill, across different regions of the country. The reports also show achievement differences according to sex, size and type of community, parental education and racial background.

Between the extremes of a classroom unit test and the NAEP are methods for every other kind of evaluation. This chapter considers evaluations of a very common type: those using statistical techniques based on test scores.

Technique and Timing: Variations on a Theme

Evaluations vary not only in technique, but also in their degree of formality. Some are carefully planned and executed. Others, "informal evaluations," are sometimes little more than a glance at test data that were gathered for some other purpose. Moreover, informal evaluation can be a euphemism for no evaluation at all.

Evaluations vary widely in their timing. Some districts evaluate continuously, virtually from week to week. Others do so on a regular basis less frequently, at intervals ranging from several months to several years. Some evaluate reasonably often, but keep to no set schedule. Still others evaluate only when they see a particular need. A small fraction of school districts never evaluate their programs.

"We use standardized test scores to evaluate programs and objectives, and board goals. They also indicate the need for evaluation by other means," explains Donald E. Schmidt, assistant superintendent for Trotwood-Madison City District in Ohio.

Others agree that testing is only a begin-

ning. "We use standardized test results to identify potential weaknesses and lead to further investigations," adds Dennis Kraft, superintendent in Aberdeen, South Dakota. And Marilyn Skinner of the Kokomo-Center Township Consolidated School Corporation in Indiana reiterates that concept: "Low standardized test scores in a particular class or building alert us to probe more deeply for reasons. System-wide low scores indicate we need to re-evaluate material, strategies, emphasis, etc."

Who conducts the evaluations? In 75 percent of the districts responding to the AASA survey, district personnel do all the work. Fewer than 2 percent contract the entire job to outsiders, but 21 percent use outside contractors to assist district personnel at least occasionally. Within the district, a variety of people might be involved. Larger districts ordinarily have special staffs or departments for evaluation. The task may be shared by curriculum staff and building personnel in the smaller districts.

The survey results suggest that tests of all kinds are by far the preferred evaluation device, with nationally distributed standardized tests the most popular. On a separate survey item, 98 percent of the districts report administering standardized tests, and 90 percent of those use the results in program evaluation.

Significant Differences: What the Results Mean

The process seems simple enough, at least in principle. Administer a test, compute the average scores and compare the averages to a standard. In recurring evaluations, as when a district evaluates routinely on a set schedule, the comparison will usually be with averages from previous studies. Often the aim is to compare two programs—for example, an ongoing program and a new one proposed to replace it. In that event, the appropriate comparison is between students in the two programs. In either case, it is usually possible to make comparisons as well with state and national averages, or "norms." (See Glossary.) Sometimes comparison with the norms is the only outcome of interest.

In practice, it is often helpful to administer tests to two groups of students before and after their exposure to different programs—pre-tests and post-tests. The important indicator then becomes the change in each student's score. If the average gain in Program A is bigger than the average gain in Program B, this is evidence that Program A is more effective. There are special methodological problems in pre- and post-testing, however.

When comparing two programs (or years), a simple check of the averages is not usually enough. Rarely will the average scores undergoing comparison come out exactly the same. Yet a small difference between the averages may not be significant. There are always some fluctuations due to chance—not a sound basis for making program decisions. It is necessary to separate the small chance variations from the other variations that might mean something. The key question—the question basic to all comparisons of this kind—is how big the differences must be in order to be considered significant.

Statistical methods can answer this question. An evaluator can dump several thousand individual test scores into a computer and, a few minutes later, be told whether their averages (or average gains, in a pre/post experiment) differ significantly. More exactly, the computer will specify how unlikely it is that such differences would arise by chance. If that possibility is very low, then the differences are considered significant. These numbers are ordinarily expressed as "confidence intervals" or "levels of significance." (See Glossary.) They can help to make an informed decision on whether this year's math performance is better than last year's, or whether the new reading program works better than the old one.

Eliminating Extraneous Factors: Testing Programs not Children

The major problem in educational program evaluations using comparison methods is to ensure that any differences in test scores are due only to differences in the programs, not in the children.

Suppose that a comparsion of two programs (or of two years) shows a clearly significant difference, far greater than chance is likely to provide. Is it safe to conclude that the higher-

scoring program (or year) reflects higher achievement? Unfortunately, this conclusion is **not** justified—not without a good deal of further information.

The entire process of comparison, and the statistical methods it uses, have a number of built-in assumptions. The comparisons are valid only if all the assumptions are true, or nearly so. Otherwise, no matter how big the difference in average scores, the comparison may well be meaningless.

In the context of program evaluation, five major conditions must be met.

1. **Good tests.** A comparison based on test scores is only as meaningful as the scores themselves. Accordingly, the tests that provide those scores must be well suited to the needs of the particular evaluation. Chapter 3 discusses what administrators should consider in choosing a test.
2. **Similar groups.** A program evaluation must measure students' performance—not to assess their individual skills, but as a means toward determining the program's effectiveness. Evaluating individual achievement is a fundamentally different task. There, sensitivity to individual children's abilities is important. In trying to draw conclusions about a program, however, any systematic differences in abilities become a source of error. The effect of such differences must be kept to a minimum.

Indeed, this is why comparison studies are so popular. By comparing two groups and looking at the difference, it is possible (in principle) to balance off the abilities in the two groups. Then, only the differences between the two programs (or years) ought to emerge. But there is a catch: the approach works only if the two groups are very similar, except for their programs.

Finding similar groups for this kind of experiment has long been a problem in the social sciences. It is much easier in fields like pharmacology and physiology, which make extensive use of the same approach in animal experiments.

Children, however, do not come in numbered genetic strains. (There have been some fascinating comparison experiments in psychology using identical twins.) Children's backgrounds are highly variegated. The problem is to divide a heterogeneous collection of students into two reasonably similar groups. There are certain routine ways of doing this, none of them entirely satisfactory:

- **Random selection**—The children are assigned to the two groups at random. If a great many children are involved, chances are the two groups will be reasonably similar. The smaller the initial pool, however, the greater the chance that systematic differences will emerge. (One fast method is to split the group by even and odd birthdays.)
- **Criterion selection**—This method uses prior results, usually test scores, to choose groups for a particular purpose. In a comparison of two math programs, for example, the groups might be chosen for their similar math scores in some previous test. The previous test must be carefully chosen, however. A test in rote arithmetic may not give equivalent groups for evaluating a program in problem solving. Ideally the test should measure all the characteristics that are important to both programs. In many cases, the pre-test will serve to divide the groups.
- **Self selection**—In this method, students themselves choose one program or the other. From a technical point of view, it is usually undesirable. Chances are the two groups will be systematically different in some respect that bears on the subject area, and so will not be equivalent at all. But often there are reasons in politics or simple fairness for letting students make their own decisions. In that event, the evaluator can make certain mathematical corrections to the results, but there will always be some doubt about the real outcome.

3. **Similar treatments.** The experiment depends on the groups being similar in all important respects, except for their programs. Thus, not only must the groups be alike in themselves, but they must be similarly treated except as the programs

require otherwise. Three examples that come up relatively often:

- Classroom time in the two programs should be the same, unless time itself is one of the issues being studied. Likewise, homework assignments should be comparable.
- Teacher enthusiasm is an important variable. A teacher assigned to one program over the other against his wishes is likely to pull down the results. Where possible, it is better to let teachers self-select in hopes of equating this factor. Teachers are not usually a large enough group for random assignment to be effective.
- Conversely, some teachers want very much to see their program succeed, and so may give their students too much help in preparing for the post-test. Sometimes this is unintended, but it can bias results nonetheless.

A host of environmental factors can also affect the evaluation results, if they are appreciably different between the two programs. These factors may include the programs' times of day, whether the various classrooms are comfortable and quiet and so forth. The environmental factors here are generally similar to those that affect test performance.

Another factor is more difficult to eliminate. If students see one of the programs un-

Help from NAEP

The National Assessment of Educational Progress (NAEP), in the words of one spokesman, "evaluates the nation." Every year it tests a representative sample made up of between 70,000 to 100,000 students nationwide. The students are all in one of three age groups: 9, 13 or 17 years old. On occasion NAEP also tests "young adults," aged between 26 and 35. Testing is in eight subject areas:

- mathematics,
- reading/literature,
- writing,
- social studies/citizenship,
- science,
- art,
- music,
- career and occupational development.

Every four years, NAEP issues reports on each major subject area that break down achievement by geographical region, size and type of community, sex, race and level of parents' education. The reports are very specific about the objectives tested, and include many examples of the actual items used. Despite the overwhelming size, these are in many respects the model of a good evaluation report.

For its access to students, NAEP requires the cooperation of local school districts nationwide. In return it offers certain kinds of help to all districts which ask, particularly in materials for information, training and planning. NAEP does not, however, release the scores of individual students in the district. In making up its statistical samples, NAEP will often search a particular district for students with particular characteristics. Thus, while the overall sample is closely representative nationwide, it is not at all representative within the one district. NAEP accordingly feels that release of local results would be misleading.

Certain of NAEP's materials can be useful to local districts in their own program evaluations. NAEP makes the following available:

1. Objectives used by NAEP in constructing its own tests—often useful for making local tests as well.
2. "Released items": test items used in the past and now retired.
3. Findings, item by item (except for the breakdowns mentioned above, which are aggregated).

Districts are free to use released items

der test as "special" in some way—if it is new, has fewer students in it, involves special equipment or the like—then its results will likely be artificially high. The mere fact of being in a special program becomes a motivating influence. Called the Hawthorne Effect, this influence was originally discovered in an industrial setting, but it affects educational evaluation as well. The extra motivation, unfortunately, disappears once the program becomes routine.

4. **Similar testing conditions.** It is easy to ensure that the physical testing conditions are the same for both groups: simply have them take the test together. And, if more than one form of the test is involved, it is not difficult to arrange the same mixture of forms for both groups.

A more subtle form of testing difference, however, can upset an evaluation badly. Even when both programs teach the same content and the test examines that content properly, the test can still skew results in favor of one group or the other.

To take an easy example, suppose one of the programs being compared uses conventional classroom instruction. The other, say, is a "mastery" program that divides the content into small pieces and tests the student on each piece before the student may proceed to

as part of their own tests. Ordinarily it is not practical to construct a complete test from released items; NAEP recommends at most a 50/50 mix of their items and those developed locally.

Many districts are interested in how their own students and programs compare with the NAEP sample. Hence, some reconstruct NAEP-type tests for local use, using NAEP items in part, and then compare the results with those published by NAEP. The outcome is much more detailed than the usual publisher's norms and is correspondingly more helpful. It is possible to make local/national comparisons not only against the test as a whole, but against individual items.

NAEP has no objection to this practice, but the directors stress that an exact replication of their testing procedures is impossible for most local districts. Thus, they feel that only very general, "ballpark" comparisons are meaningful. They also have some specific words of caution:

1. Make sure that the objectives on which the NAEP items are based match the objectives used locally. The NAEP objectives tend to be broad in scope, so as to embrace a wide range of instructional programs. Local objectives may be much more specific.
2. Inspect the items one at a time to make sure the content of each was actually taught in the local schools and received sufficient priority.
3. Keep in mind that NAEP chooses students by age, not grade. The group of 13-year-olds, for example, includes students who differ in age by almost a full year—and they might be in several different grades. The time of year at which local testing occurs can also lead to results different from NAEP's.
4. Keep in mind also that NAEP reports come out between 12 and 18 months after testing is complete. Where student populations are changing rapidly, this time lag can make a difference.
5. For the best possible comparison, it is necessary to use testing procedures similar to NAEP's. Those are expensive, however. NAEP sends its own testing personnel into the schools to ensure uniformity. Students not only read the questions on paper, but also hear them read aloud on tape. Many of the items require hand scoring, a possible source of error unless the graders use NAEP's criteria throughout.

the next. Chances are the mastery program will show up better on a multiple choice test, whether or not it is more effective pedagogically. There are three reasons.

First, the testing situation will be more routine and less threatening to the mastery students, because they have gone through it more often. Second, items on the post-test may well be similar to items students have seen on the mastery tests; the multiple choice format, after all, offers only a limited number of ways to test any particular content. The mastery students will have seen those items, and will have had feedback on their mistakes. Third, a mastery program and a multiple choice test must divide up the content in much the same way. A conventional student may be hampered by thinking too holistically. The mastery students, however, will be accustomed to looking at the content in one small particle at a time—a perspective that will serve them well on the post-test.

5. **Good distributions.** This is a rather technical point, but an important one. The most widely used statistical methods for finding significant differences require that the test scores follow a "normal distribution"—the bell-shaped curve. If the scores depart too much from a normal curve, the results of the statistical analysis are not valid.

There are mathematical tests available to determine whether scores follow the normal distribution closely enough, and a conscientious evaluator will apply them. This step is advisable even when the test manual advertises normally distributed scores. The test's publisher can ensure only that the scores of students nationwide follow a normal distribution—not that scores from this particular evaluation do.

If the scores are not too far away from the normal distribution, mathematical corrections are possible. Otherwise, it may be necessary to fall back on other methods, called "distribution-free statistics," that impose no requirements on the patterns of the scores. But these methods are less sensitive to small significant differences. A distribution-free test may report no significant differences whereas another method would find some. More important, a bad distribution may signal trouble in the evaluation.

Pre- and Post-Testing: Problems in the Experiment

Many evaluators are particularly fond of the pre-test/post-test procedure. This model tests students before and after they take part in their respective programs. The evaluation uses not the scores themselves, but the difference in score that each student shows from the pretest to the post-test. These, called "gain scores," are averaged and compared. The program showing higher gains is considered the more successful program.

A similar procedure is often used in comparing students' scores from year to year. Even if only one program is involved, the evaluator may look to individual gains over the year as an indication of the program's success.

However, the pre/post model has problems of its own that sometimes outweigh it advantages.

Reliability. Gain scores are sometimes considered much less reliable than the underlying test scores. "Less reliable," in its technical sense, means that students are likely to show different gain scores on different occasions. That being so, any one gain score must be viewed with suspicion. (See Chapter 3.)

The reason is not hard to see. Every test score has a certain margin of error. That margin may be small enough when compared to the test score. When one subtracts two scores, the error remains—but now one must compare that error to the number planned to use: the difference between the scores. The difference is much smaller than the scores, and so the same error become relatively larger. (Subtracting two scores changes the error too, but that is less important here.)

Initial differences. Any initial lack of similarity between the two groups—especially differences in their pre-test scores—will probably affect the final results. Merely subtracting the scores does *not* cancel out the initial differences. Even if the programs were exactly equivalent—even if they were the same program—a difference in the pre-test scores causes a difference in the subtracted gain scores as well.

Which group will show higher gain scores? Much of the time, it will be those who scored lower in the pre-test. This is usually the case when the pre-test and post-test scores have similar distributions. It is a surprising result. An evaluator might expect the higher-scoring students, who may be more capable in some respects, to learn more in the program. On the other hand, the low pre-test scorers have, quite literally, more room for improvement.

Under certain conditions, the high pre-test scorers will show higher gain scores. These conditions are most likely to arise when the test scores are reported in grade equivalents.

What scores to subtract? Test scores are reported in many forms. Some should not be subtracted to compute gain scores. Percentiles, for example, are perhaps the most widely used form of reporting. But the differences between percentiles cannot properly be used to infer educational gains. Indeed, no score should be used in this way if it directly represents a student's rank, rather than actual achievement.

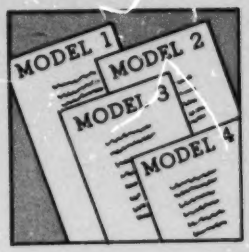
Grade equivalent scores, in particular, can present a host of problems. They remain very popular nonetheless, especially for recurring evaluations. They seem to show at a glance whether students are making enough progress from year to year: a gain of one grade equivalent for each year in school.

That standard is only appropriate for students near the middle of the range, however. Above-average students will ordinarily make greater gains, and below-average, lesser gains. Expectations should be adjusted accordingly. This adjustment can be difficult in a program evaluation, which may involve students at several different grade equivalent levels.

Time between tests. Programs of short duration—a few months—ordinarily require special pre-tests and post-tests. But where a school district tests once a year in any event, it will naturally want to evaluate its year-long programs with the results. The program only runs from fall to spring, however. In principle, testing at both ends of the school year (fall/spring) would give a more direct indication of gains due to the program.

Whether annual testing (fall/fall or spring/spring) gives the same results as fall/spring testing depends on what happens to student achievement over the summer. This assumption is presently a matter of controversy. For norming purposes, test publishers have traditionally assumed that the three summer months result in the same gains as would one month in school. This makes the calendar year effectively ten school months long. Though there is no apparent support for this assumption, it does make the mathematics easy.

Recent studies show conflicting results on the question. Some find declines during the summer, especially in math. The declines can



Popularity of Evaluation Methods

The AASA survey for this publication asked how often districts use each kind of evaluation technique.

	Usually	Occasionally
National standardized test	78%	16%
Teacher-made test	56%	18%
Reports by principals	54%	34%
Inspection of student work	52%	31%
Pre- and post-tests	52%	36%
In-classroom observation	50%	34%
Minimum competency test*	44%	13%
Oral report by teachers	41%	38%
Written report by teachers	39%	42%
District-written test	28%	24%
State-written test	26%	12%
Parent interviews**	18%	55%
National criterion-referenced test	17%	24%
Interviews with students	9%	38%
In-classroom videotape	6%	22%

* These figures represent all districts, including the 48 percent that report no mandated competency testing program. According to a separate item, of those districts that are subject to competency test mandates, 84 percent use the results in program evaluation.

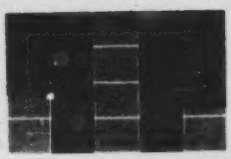
** Several districts volunteered that they send questionnaires to parents and that they take the results into account in program evaluation. There is no accurate count on this practice.

be a good fraction of the gains from the preceding year. Other studies, however, show gains over the summer, and some indicate little change in either direction. Results typically vary with subject matter and grade level.

For program evaluation, fall/spring testing is the safest approach, though often expensive and inconvenient. Fall/spring testing also helps to minimize another problem: changing student populations. A student who leaves the program partway through, or who comes in after the pre-test, will upset the results unless that student's scores are dropped from the study. But the dropping of too many scores can disturb groups that were carefully equated by pre-test. This kind of attrition will be less severe if the two tests are separated by the least possible amount of time.

Chapter 3

Choosing a Test: The Right One for the Purpose



Choice of a test is often the most important single decision in planning an evaluation. To do its job, the test must meet two requirements. First, it must be a good test in its own right. Second, it must suit the needs of the evaluation. These requirements are not always the same. A test that is excellent for one purpose can give misleading results when used for something different.

What makes a test a good one? It must be both reliable and valid. A test is reliable if it gives the same results consistently. It is valid if it measures what it claims to measure.

The words "reliable" and "valid" have technical meanings in testing different from what they mean in ordinary English. A test can be "reliable," but still unreliable for certain uses; it can be "valid," yet still give meaningless results. The double meanings of these words allow test publishers to make some perfectly truthful claims that are, at the same time, misleading to unsophisticated users.

Test publishers make available a lot of data on their products' reliability and validity. Another word of caution, however. There are methods to boost artificially a test's reliability. And it is possible to make almost any test look valid, even if it is not. The best assurance that the data mean what they say is a reputable publisher.

Some evaluators look no further than reliability and validity, largely because these factors are so easy for the user to ascertain: one need only look them up in the test's manual. Properly, however, reliability and validity are only starting points. They are enough to disqualify a test, but not nearly enough to qualify one.

It is very important to go beyond the publisher's data and to examine the test itself. Often this approach is the only way to ensure that the test suits the purpose at hand. But here, too, many evaluators stop at the manual. They need a math test for third graders; the title of the test reads "Level 3 Math"; and that seems to be enough.

Looking at just the title is not nearly enough. There are many approaches to teaching math, many possible priorities and emphases. The word "math" in the title says very little about the test. Some publishers provide lists of the objectives tested, but these too can be misleading. Only the questions themselves can tell what the test is really about.

Reliability of Results: A Crucial Characteristic

A test's reliability refers to how consistent its results are and to how much error is likely to be found in any score. This characteristic of a test is easiest for the publisher to determine.

Every test result is in error to some degree. Some tests have a higher degree of error; some, a lower degree. Think of a cloth tape measure that stretches a little. The stretchier the tape, the more two measurements will disagree. And also, the stretchier the tape, the more error there will be in a single measurement. Tests too differ in how much they "stretch." In the same way, their likely degree of error varies from one test to another.

High reliability in a test means a high consistency, a low degree of "stretch." High reliability also means a lower margin of error in each score. High reliability is thus desirable. But note: the kind of test error that reliability reflects is only one of the many ways a test score can be wrong. Thus, even a highly reliable test can still give erroneous results.

Finding the Reliability: Testing the Test

Ordinarily a publisher estimates the reliability of a test toward the end of its development, when it is nearing its final form. Here are three main methods of assessing reliability:

Test/retest: This way of finding a test's reliability is analogous to measuring twice with the tape measure. A population of students is carefully chosen to represent students who will ultimately use the test, after it is on the market. Each of these students takes the test twice. The publisher compares the two scores, student by student. The closer the scores are, on the average, the better the reliability.

There is a problem, however. Students learn. What a student learns while taking the test the first time can affect the score the second time. The publisher may apply a correction factor to compensate or turn to one of the other methods.

Equivalent forms: Many tests are published in "equivalent forms," two or more versions of the test that examine much the same skills in essentially the same way. All things being equal—except for test error—a particular student should obtain the same score on any of the equivalent forms. To measure reliability, then, a representative group of students takes two of the equivalent forms. Due to test error, their scores will not be exactly the same. The difference in scores, computed over many students, gives an indication of the test's reliability. The lower the difference, on the average, the higher the reliability.

This method, too, has its limitations. For one thing, it is hard to be sure that two forms of a test are in fact exactly equivalent. And even though using different forms, a student may learn something on the first test that helps him or her on the second.

Split-half analysis: If the test is moderately long, it is sometimes possible to split the test into two tests that are roughly equivalent. The simplest way—it does not always work—is to regard the even-numbered questions as making up one test, and the odd-numbered questions as making up another. More sophisticated methods are often necessary. Each student in the sample group takes the test at one sitting. But the results are analyzed as if each student had instead taken two half-sized tests. The scores for the two halves are figured separately and compared. If the scores from the two halves are very similar, on the average, then the test has high reliability.

Most methods for determining the reliability have students taking a test twice or students taking two tests. But most evaluations have students taking one test one time. Reliability is important because it gives an indication of the likely error in every score from the test.

The analysis of a reliability study is rather complicated and is usually carried out using a computer. The result is a single number, called the "reliability coefficient." It wraps up in a couple of digits almost everything there is to say about the test's reliability and error.

The reliability coefficient is a decimal number somewhere between 0.00 and 1.00—the higher the better. A test whose reliability were zero would have no reliability at all. At the other extreme, if the reliability were 1.00, the test would have perfect reliability. It would have no test error of this kind at all.

Most of the widely-used achievement tests have reliability coefficients ranging roughly from 0.90 to 0.94. Some aptitude tests in common use go as low as 0.80. A reliability coefficient much below 0.90 should be cause for some concern. It says, in effect, that the mar-

gin of error is moderately wide. For coefficients much under 0.80, the margin of error becomes so broad as to render results from the test all but useless.

The designer of a new test knows that prospective users will look at the reliability coefficients carefully. Accordingly, the designers do what they can to keep the coefficient high.

But artificial methods of boosting the reliability coefficient do not necessarily help create a better instrument. For example, other things being equal, a longer test will have a higher coefficient than a shorter one. Once the test format is established, adding more items of the same kind is usually not difficult. But examining the same content twice, or four times, may not make the results any more meaningful.

Another way to raise the coefficient is to shorten the time limits until most children cannot finish the test. But doing so changes the nature of the test significantly. It becomes less a measure of the child's academic skills, and more a measure of his or her speed. Unless evaluators are genuinely interested in speed, and few are, the end result is poorer for evaluation purposes, despite its better reliability. And some learners react badly to the stress of a tightly timed test.

The important question is not the reliability coefficient itself, but what the coefficient says about test error for each child. The coefficient and the likely test error are connected through a set of formulas. Few evaluators bother to make the calculation. Instead, most merely use the coefficient itself as one factor in deciding which test to use. It is worthwhile, all the same, to know just how large the test error can be for even a relatively high coefficient.

Test Error: Avoiding Erroneous Results

Test error is less a problem in evaluations that average the scores from many students. Then, the test errors tend to cancel out. The average score should not be too different from the average "true score."

Again, however, the kind of test error discussed here, the kind associated with reliability, is only one way a test can give erroneous results. Some errors do not average out.

Averaging scores will not always cancel even "reliability error." It works only if the students being averaged are a roughly homogeneous group. Averaging the scores from student populations that are very different not only gives rise to test error, but is also likely to give misleading results for other reasons.

Whenever the average involves fewer than about 30 students, enough test error remains to cause serious problems. It is probably safe to average students in one classroom, but a group any smaller will show significant error. And the error becomes especially important when individual students' results play any part in the evaluation, as they usually should. A high reliability coefficient is the only protection available, but even that is not enough. Even the best instruments available leave enough test error to raise serious doubts about the meaning of any one score for any one student.

Verifying the Validity: Another Test for the Test

Along with reliability, the other test characteristic that figures heavily in publishers' manuals is "validity." Of the two, validity is by far the more important. It is also much harder to determine.

A test is valid if it measures what it claims to measure. It is not valid if it measures something else instead, or worse, if it measures nothing at all. Notice that validity depends both on the test itself and on the purpose of the test. A test that is valid for one use may be completely invalid for something else.

Here is an example, somewhat contrived to make the point more clearly. Suppose a certain test is said to measure mathematical problem solving skills. The problems on the test are mathematically quite simple. But to solve each problem, the child must first read it. Due to careless design of this hypothetical test, the reading level is quite difficult for the intended grade. What, then, is this test measuring? Under the circumstances, not mathematical problem solving. It is more likely to function as a reading test. That is, a child's score will depend more on his or her reading skills than on math abilities.

As a math test, then, this is not a valid instrument. It does not measure math skills. Might it be valid as a reading test? Perhaps; but only if the math is so easy that virtually any child can do it. Otherwise the test is measuring some unknown combination of math and reading. It would then be valid for neither math nor reading, and it would be useless as a test of either one.

Striving for Perfection: One Test for One Skill

An ideally valid test would measure only one set of skills, those specified in the test manual. This cannot be fully achieved in practice, however. Other skills will always affect the results. For example, nearly every test measures reading skills, because a student who cannot read the questions will always obtain a low score. Every test that uses a separate answer sheet also measures the student's ability to keep track of which place on the answer sheet goes with which question. (This is not always an easy task, especially for a student who chooses to skip a question here and there.) And there is a set of skills, in test-taking itself that affects every student's score.

There are about a dozen different techniques for estimating validity. For most tests, not all will apply. Unlike determining reliability, each method of finding validity tells something very different about the test. A few methods do give "validity coefficients," which are roughly analogous to reliability coefficients. But the validity coefficients from different methods are not comparable. And most methods of assessing validity do not give numerical results at all, which makes comparisons even more difficult. All the methods seek to answer the same basic question—Does this test measure what it is supposed to?—but they deal with the question in markedly different ways. Worse, no one method can ever provide a clear answer. Even a combination of methods will often leave the question unanswered.

Note, however, that the validity of a test cannot be determined from the manual alone. Validity depends not only on the test itself, but also on the intended use of the test. No test is valid in itself. Purpose is always a key issue in validity. No matter how reassuring the manual is, the ultimate decision on validity—on whether the test is a good one for the purpose—can be made only by the users, the people who know what they expect to learn from the test.

Questions about Tests

From George S. Vrabel, director of elementary education, Mishiwaka, Indiana:

Working on the assumption that your local district is using standardized testing in some form, four questions need to be explored and analyzed. These questions are intended as a starting point for later practical applications of standardized testing. If the response is negative to any of the questions, your standardized testing program may not be utilized properly. . . .

1. Does your district have stated objectives for the use of standardized testing?
2. Has your standardized testing program been evaluated by the administration and teachers within the last three years?
3. Is the standardized testing program meeting the needs for student evaluation?
4. Is the standardized testing program meeting the needs for improvement of instruction?

Assessing Validity: Methods for Evaluators

Test publishers have developed a wide variety of methods for determining a test's validity. Each is appropriate only for certain kinds of tests. Ordinarily the testing manual will present several methods—as many as possible—leaving it to the evaluator to figure out which ones make the most sense.

The names for the methods suggest different "kinds" of validity. All, however, address the same underlying question: whether the test measures what it should.

Face Validity: This is by far the simplest. A test is "face valid"—literally, valid on its face—if it merely looks like the right kind of test. This makes some sense in the case of rote mathematics. A test that looks like a test of subtraction probably is one. On the other hand, a multiple choice test that purports to measure writing skills will have a lower degree of face validity.

Content Validity: Very important in school achievement testing, content validity rests on an analysis of the "content" the test is supposed to examine. It uses the same list of instructional objectives the test authors used in constructing the test. (The list, or a revised version of it, often appears in the testing manual.)

An expert analyzes each item on the test to be sure that the content it measures is among the test's stated objectives. Preferably each item will test only one of these objectives. The expert then makes sure that all of the objectives are represented in the test, and that each objective receives sufficient attention. A test that meets these requirements is said to be "content valid."

Virtually every achievement test on the market claims to have content validity. Two cautions are necessary, however. First, content validity rests solely on expert opinion. Second, content validity is keyed to a specific list of objectives.

Concurrent Validity: This method finds occasional use in achievement testing. It compares results from the test in question to the results from some other measure given concurrently; hence the name.

Concurrent validity is often used for validating IQ tests, since there is no possible measure of IQ apart from some other IQ test.

Concurrent validity can yield a "validity coefficient." Like reliability coefficients, these numbers range between 0.00 and 1.00. But they tend to be far lower than reliability coefficients. Even the best typically fall well under 0.60, as compared with over 0.90 in the case of reliability. The low coefficients suggest, once again, that test scores depend on factors other than the skills supposedly being tested.

Concurrent validity rests on two assumptions, both of them important. First, the children taking part in the study must be truly representative of the children who will eventually use the test. Second, the measure used for comparison must be a good one.

Predictive Validity: Not much used in achievement testing, this technique finds its greatest application with aptitude tests. Like concurrent validity, it compares the test in question with some other measure. In predictive validation, however, the comparison measure is given some months or years after the test.

Claims of predictive validity ought to be viewed cautiously, in view of the generally low numbers. When a test is used for prediction, the validity coefficients are essential to an informed judgment about the test's worth. A testing manual that claims predictive validity, but does not show the coefficients, deserves great skepticism.

Construct Validity: Until the rise of the minimum competency movement, this method played little part in educational testing. It applies to tests that examine some characteristic of the student that is not observable directly, but which is assumed to affect behavior. Examples of such characteristics are intelligence and anxiety.

None of this had much to do with educational evaluation until recently, with the rise of the minimum competency movement. In certain states, the legislature required tests of "functional literacy" and the like. Functional literacy is hard to define precisely, though we can give many examples of it. Functional literacy is a construct, just as intelligence and anxiety are. Tests of functional literacy are subject to construct validation.

Many states and districts that mandate minimum competency avoid the construct problem simply by avoiding references to functional literacy or anything like it. Instead, they specify detailed lists of instructional objectives. The appropriate method then becomes content validation, which is a simpler procedure.

The remaining two forms of validation also arose very recently, and also in response to the competency testing movement.

These types differ in emphasis from those previously mentioned. The types already discussed evaluate a test in terms of the purpose

it is meant to serve. The two forms which follow bring in a third element: the student. They ask whether the test serves its intended purpose in terms of each student as an individual.

Curricular Validity: For a test to have curricular validity, the content it examines must have been part of the student's curriculum. A test might meet several of the validation criteria listed above and still fail here. Curricular validity does not validate the test itself. Rather, it validates administering the test to a particular student.

From an evaluation perspective, a lack of curricular validity will render test results meaningless. There is little to be learned from measuring skills that do not appear in the curriculum. In the minimum competency environment, however, the issues become far more serious. It is indeed unfair to penalize a student for not learning material the school never tried to teach.

In most cases, establishing curricular validity is not difficult. It involves merely checking the list of objectives for the test—the same list used for content validation—against the curriculum. The test has curricular validity if all of its objectives were sufficiently addressed in the curriculum.

Instructional Validity: This form of validation takes curricular validity one step further. It asks not only whether the tested material was part of the curriculum, but whether it was actually taught to each individual student. The arguments are the same as for curricular validity: meaningful results in evaluation and fairness in competency testing.

Item Analysis: Questioning Each Question

Most tests undergo extensive tryouts before publication. Ordinarily the tryouts, especially the later ones, involve groups of students chosen to represent the test's ultimate users in terms of age, sex, grade level, geographic location, race, socioeconomic status and so on. Late in this process, when the test is nearly completed, its reliability coefficient is computed, along with coefficients for concurrent or predictive validity where these apply. But earlier tryouts can result in significant changes to the test itself. One step in the process called "item analysis" has particular influence on the final composition of the test.

Item analysis requires keeping track of every student's answer to every item. Imagine a huge table of test scores: individual students listed down the side, item numbers across the top, and "right" or "wrong" in each little square. (In practice the table exists only in a computer.) When the test is in actual use we shall be interested mostly in counting across the table, to find how many items each student answered correctly. Item analysis counts down instead, to find out how many students answered each item correctly. This information is used to drop some items from the trial test and to revise others.

For example, an item will not survive if nearly all students answer it correctly, or if almost none do. One goal in test design is to distinguish among students with as few items as possible. An item that nearly everyone answers the same way does not serve that goal.

Some items may pull down the test's reliability overall. These are items to which many children give inconsistent responses on retest, or between forms. They too are likely to be dropped. Similarly, items that pull down the validity coefficients may be removed, if they can be identified.

The analysis examines not only how many students got each item right, but also which students got it right—and how well these students did on the test overall.

Item analysis tends to make the test more "efficient." By eliminating those items that clutter the scores with inconsistencies, the designer arrives at a test that gives more meaningful results with fewer items. But the process of dropping items on statistical grounds can also change the overall thrust and emphasis of the test.

Eliminating Test Bias: Protecting Against Problems

When a test is less valid for some groups of students than for others, the test is biased. Putting it another way, a test is biased if a student's score depends not only on his or her academic ability, but also on race, sex, ethnic origin or some other group factor. None of the validity procedures described in the preceding

How to Inspect a Test

In weighing a test's validity, it is almost always helpful to go beyond the manual and inspect the test items themselves. These are some of the relevant questions:

1. Does the test look like the right kind of test?
2. Does the test cover all the content the evaluation requires? Are the emphases appropriate?
3. Is there content in the test that is not part of the evaluation? If so, is it enough to upset the results?
4. Is the test's terminology the same that students have been exposed to in the classroom?
5. Is the reading level appropriate?
6. Is the format clear? The instructions? The answer sheet? Will the students understand what is expected of them throughout?
7. Are the questions reasonably unambiguous? (Many are not. Try to read questions from the student's perspective.)
8. Are all the right answers in fact right? All the wrong answers wrong?
9. Will any be offensive or confusing to any of the ethnic minorities in the student population? To either sex? To handicapped students, one-parent children or others in special circumstances?

pages offer any protection against bias.

Two forms of test bias are all but unavoidable, and for that reason are not usually treated as bias at all. Some educators say that most tests now in use in this country discriminate against children who are not fluent in English or against children who have handicaps that limit what they can do in the testing room. With care, in some testing situations, these two problems can be held to a minimum. But they can rarely be eliminated altogether.

Should handicapped and non-English-speaking children be included in program evaluations? The answer depends, once again, on the circumstances of the evaluation. If the school or district serves substantial numbers of children who are handicapped or who do not handle English fluently, leaving them out of the evaluation will yield distorted results. On the other hand, if they are few in number, their effects on the final result will be small.

One resolution is to include these children in the evaluation, but to make a special inspection of their test scores afterward. If their scores are much below average, it is important to distinguish between two kinds of reasons: (1) because the program under study is not serving these groups of children well; or (2) because the test itself did not reflect their capabilities. It is usually necessary to supplement the test with other methods of evaluation, especially if the program undergoing evaluation is particularly intended to meet these children's needs.

It is possible to eliminate other forms of test bias, or at least to reduce them greatly, through careful test design. As part of the test development process, many publishers now include a special review by women and by members of the larger minority groups. Their task is to identify and weed out potentially biased items. As a rule such items can be rewritten to eliminate the bias.

Striving for the Ideal: Allowing for Varied Experience

Some test bias is all but inevitable. Any test requires much more than academic skills. All but a few tests use language. Language depends on words and phrases carrying the same meanings for the child as they do for the test author. This in turn requires that the child has shared certain experiences with the author—for language, in the end, is a way of coding experience. Often values come into play as well, even in subject areas that ostensibly have nothing to do with values. For example, authors of math tests like to frame questions

in terms of family situations, believing that such questions will be more meaningful to children. But it is in the context of the family that minority cultural values may differ most sharply from those of the majority.

Children's home environments differ. Their experiences in growing up are not the same. The values stressed by their families vary. It is hardly surprising, then, that children from different cultural groups may understand the same test question differently. This observation does not, of course, relieve the test author—or the evaluator—of responsibility for test bias. The only bias ever justifiable is the irreducible minimum that no amount of revision or redesign can eliminate. And even that must be taken into account in interpreting scores.

Some publishers raise this defense against charges of bias: that minority children, in the appropriate proportions, were included in the test's standardization sample.

As part of his or her professional responsibility, an evaluator should plan to inspect every test used, for possible bias. There are three kinds of problems to watch for. Each applies both to minority children and to children of either sex: (1) elements that may be harder for some children than others to understand; (2) elements that may be offensive to some children; and (3) elements that may cause some children needless anxiety.

Whether a particular test item presents these problems or not will depend to some extent on the local student population. The concerns of a district serving many Indochinese children, for example, will be different from those of a district whose minority children are mostly Hispanic.

It can be helpful to have a few adults from the local minority communities examine a test in search of items that may give their children special trouble. Such a group must be chosen carefully, if it is to function well. A word of caution, however. Security-minded publishers often require schools to sign agreements that limit access to their tests. A local review for bias may require the publisher's permission. Should the publisher refuse, that may be sufficient reason to try another.

What to Look For: Examples of Bias

The following are examples of elements that can lead to bias in tests:

- Use of words that may be unfamiliar to some cultural and regional groups;
- Illustrations of objects unlikely to be found in some homes;
- Assumed familiarity with properties of objects not equally available to all children;
- Assumptions that children have money of their own to spend;
- References to holidays, traditions, etc. that have a religious or cultural origin;
- Sexual stereotyping of male/female roles in employment, housekeeping, child care, home repairs, etc., or consistent reverse stereotyping;
- References to sports, games, pastimes engaged in by one sex substantially more than the other;
- Consistent depiction of nuclear two-parent families.

Administrative Concerns: Time, People and Money

This chapter has examined the educational and psychometric factors affecting the choice of a test. These factors help to answer two questions: (1) Will this test serve the purposes of the evaluation? (2) Can the results of this test be trusted?

In practical terms, other factors become just as important in the selection. These are the administrative concerns involving time, people and money. In comparison with the factors just discussed, the administrative information is more readily accessible. Indeed, most publishers are eager to provide it.

Before settling on a particular test, the evaluator should have firm answers to the following questions:

- How much will the test cost? (per student; per classroom; overall)
- What does the price include? (test booklets and answer sheets; scoring manual; technical manual; scoring services; individual reports; summary reports)
- How long will the test take to administer?

If necessary, can it be administered in two or more sittings?

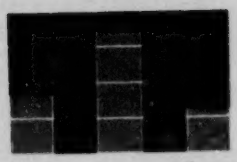
- Does administering the test require specially trained staff?
- Will students taking the test need any special equipment or materials?
- If the publisher scores the test, exactly what reports are provided? Will the publisher break down the results to meet special needs for the evaluation, if necessary? Is there an extra charge for this service? (Important: Discuss any special computations with the publisher before the test is administered.)
- Does the publisher provide an individual report suitable for sending home to parents?
- How long will the publisher take to score the test and send back the results?
- How complete are the manuals? Do they provide all needed information?
- Does the test cover what the schools are actually teaching?

Because this information is so easily available from the publisher, it is tempting to rely on it too heavily when choosing a test. Some evaluators merely "shop around" among the major publishers for price and response time, paying little attention to either the tests' technical characteristics or to exactly what the tests examine.

Unless the test asks the right questions, however, the evaluation cannot give the right answers. The results of an evaluation are no better than the test that provides them. Accordingly, the selection of a test will have an important effect on the outcomes of the evaluation. It is a major decision, and often a difficult one. Only after carefully balancing the many factors involved can an evaluator be confident of having made a good choice.

Chapter 4

Using a Test: Eliminating Environmental Influences



The purpose of a test in program evaluation is to find out how much children are learning from the program. Test results, however, can depend on factors other than the academic skills supposedly being measured—on factors other than program effectiveness. Chapter 3 pointed out a number of problems inherent in many tests.

This chapter examines circumstances beyond the test itself that can influence scores: the testing environment, test anxiety in students and test-taking skills.

Most of these elements are subject to legitimate changes. By optimizing each, it is often possible to move scores upward, sometimes substantially. Doing so with some of the factors is unremarkable. For example, defects in the testing environment that represent barriers to student's performance should be corrected whenever possible. Other factors, however, such as teaching to the test or instruction in test-taking skills, are subject to more debate. The advisability of addressing these, in each particular case, must be left to the professional judgment of the individual evaluator.

Testing Environment: The Physical Situation

Research has borne out what all of us know from common sense: children do better on tests when they are fresh, comfortable and alert.

Time of day can affect performance. Some experts advise against testing just before lunch, when students are hungry, or just after lunch when they feel grogy. First thing in the morning is not a good time either; children need time to settle down. On the average, scores will probably be highest for tests given in mid-morning or mid-afternoon. Likewise, tests given in mid-week may show higher scores (and also better attendance) than tests on Monday or Friday.

Children's sleep and food can also be important factors. It may sometimes be useful to warn parents a few days ahead of time that testing will take place. Most parents, having a strong interest in their children's scores, will help to make certain of enough sleep and a good breakfast. But sending word home may also contribute to test anxiety in the children.

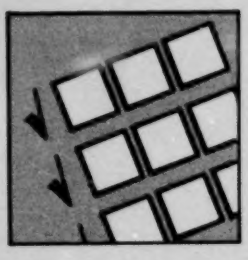
Choice of the testing room can make a difference. Many schools habitually gather children in the auditorium, library or cafeteria to administer tests. This requires fewer staff than working in the individual classrooms, and it also ensures that all of the children receive the same instructions and have the same time available. But children will feel more at ease at their own desks. A familiar environment helps to make the situation feel more routine, less "special." And particularly in auditoriums, the arrangements for writing are often inadequate. Students need enough space to spread out their materials properly. One

study has shown that students obtain higher scores writing on a smooth surface than on a bumpy one.

Conditions in the testing room have been shown to influence performance. Temperature is especially important and can become a problem when many students are grouped in a large room. A twelve-year-old gives off as much heat as a 150-watt light bulb. Eight children equal a good-sized space heater. A room that is comfortable when empty can become oppressively hot after a hundred children have occupied if for an hour or so. Similarly, good ventilation makes a big difference.

Sound level is important, and can be troublesome in cafeterias. A high level of background noise causes fatigue, even when it is steady and unobtrusive. And a large percentage of children will look up at the sound of a door opening, every single time. Such distractions should be minimized. A sign on the door is usually enough.

Students should have comfortable chairs of the right size. For younger children, cafeteria tables and chairs may lead to dangling feet—which blocks circulation—and too high



Test Administrator's Checklist Cleveland County Schools, North Carolina

Before the Testing Day:

_____ Announce the test days in advance. Prepare the students by explaining that the testing has a purpose that is important to them; i.e., to help them set goals, to help find their strengths and weaknesses, etc.
_____ Read the Cleveland County Schools *Test Administration Guide.*
_____ Read the Examiner's Manuals and take the test yourself in order to be thoroughly familiar with it—**do not coach from the test** and thus invalidate the results.
_____ Rehearse reading the directions aloud and note the time limits of each section.
_____ Plan a testing schedule with the test coordinator; pace the testing schedule so that you avoid rushed sessions or fatigue; plan for breaks.
_____ Prepare a class roster with correct birthdates of the students.
_____ Draw a model of the answer form on the chalkboard which shows
  _____ How the identification data are to be supplied.
  _____ How the circles are to be completed, densely marked.
_____ Check your supplies.
  _____ Soft #2 pencils with erasers, several per student.
  _____ The correct level of test in sufficient numbers.
  _____ The correct kind of answer form in sufficient numbers.
  _____ A group information sheet and a group information envelope for each classroom by grade.
  _____ A "Do Not Disturb" sign.
  _____ An accurate watch or clock with a second hand.
  _____ Scratch paper.
_____ Remove from bulletin boards and walls anything that will supply answers to test questions.
_____ Make arrangements so that students will not be interrupted by announcements, fire drills, visitors, etc.
_____ Decide on seating arrangements in classroom during testing.
_____ Arrange for practice session for students prior to the first test.
_____ Plan for a smooth method of distributing materials.

Just Before Testing:

_____ Prepare students for testing. Be careful not to communicate any anxiety. Try to increase their interest and willingness to do their best. The testing is designed to assist them in discovering what skills they have learned and in which areas they need more help.

a writing surface. Also for younger children, unfamiliar faces in the room should be kept to a minimum. Both of these points argue further for testing in the regular classroom.

Well before the test is to begin, make certain of an adequate supply of materials: test booklets, answer sheets, pencils and good erasers. Each child should have at least two sharpened pencils, to avoid interruptions in case one breaks. The erasers that come on pencils will sometimes not remove pencil marks well enough for the scoring machines. Pink erasers are better for the purpose.

Tests usually have instructions in the manual that teachers are to read aloud before the test, word for word. Often these instructions are badly written. Teachers should deliver the instructions as given, but can make them much clearer by reading with a lot of expression. Many teachers read this material in a flat, mechanical voice, which makes the instructions hard for students to follow. Students may also take an overly formal tone of voice to mean that the teacher is anxious, a perception that can promote further anxiety among the children.

_____ Put up "Testing—Do Not Disturb" sign and close the door.
_____ See that all desks are clear.
_____ Seat students according to predetermined plan.
_____ Check lighting and ventilation.
_____ Allow students time for restroom visits.
_____ See that each student has a library book under his/her desk in case he/she finishes a test early.
_____ Have extra pencils and scratch paper available in an accessible place.
_____ Have examiner's manual and a demonstration copy of the test open and ready.

During Testing:

_____ Distribute materials.
_____ Caution students not to begin until told to do so.
_____ Be sure that the students are equipped.
  _____ Answer forms.
  _____ Test books.
  _____ Two #2 pencils each.
  _____ Erasers.
  _____ Scratch paper.
_____ Explain the model of the answer form that is on the chalkboard.
_____ Direct the marking of all information data, if the students are going to do the marking; make sure all identifying information is marked correctly.
_____ Review the rules for using machine-scorable forms.
  _____ Marks must be made by pencil.
  _____ Marks must be dark and fill the circle completely.
  _____ Erasures must be clean.
  _____ Item numbers must be carefully followed.
  _____ There is a correct place for each test section.
  _____ Double marking is to be avoided (no two circles in the same row for an item).
_____ There should be no other marks on the answer form than the filled circle that indicates a response or helps to complete the identification data.
_____ Administer the test as directed in the examiner's manual.
  _____ Read all directions clearly and distinctly.
  _____ Follow all instructions **precisely.**
  _____ Observe all stop and start commands and time limits **strictly.**
_____ Prevent a student from looking at a neighbor's paper (keep this in mind when you make your seating arrangement).
_____ Permit no outside interruptions.
_____ Record attendance on test days on the daily report of absentees and incomplete tests to be handed in to the test coordinator.

If the test manual does not forbid it, teachers can give the instructions a second time in their own words. Likewise, unless the manual says otherwise, teachers should ask if there are any questions about the instructions. But tension in the room is often high at this point. Many children will not ask a question even if they are genuinely confused. And many teachers are too abrupt: "Any questions? Fine. Begin." The invitation for questions should be followed by a very long, encouraging pause, and then should be repeated.

In summary, the goal in setting up the testing environment is to help students feel as easy as possible. Familiarity is one important factor; physical comfort is another.

Text Anxiety: Fear of Results

Psychologists have given much attention to the effects of anxiety on performance generally. Their findings are easy to sum up: a little anxiety improves performance; too much anxiety impairs it.

Test anxiety, as the term is used here, is an anxiety about tests in particular intense enough to interfere with performance. Only recently has test anxiety become an object of study in the wake of "math anxiety," "science anxiety" and a host of other special anxieties. There is still not a great deal known about test anxiety, besides what we know from common sense.

In most cases test anxiety seems to arise not from fear of the test itself, but from fear of the results. The student is afraid of doing badly, and that fear pulls the score down. That makes the fear of doing badly even worse the next time, which leads to even worse scores—and so on, in a vicious cycle.

Test anxiety probably affects all students to some degree. But for a few, its effects are so severe as to make a significant difference in scores. The resulting low scores not only represent a source of error in an evaluation, but they also point to a good deal of suffering. Alleviating test anxiety can both help to make the evaluation more accurate and to make a few students' lives much more pleasant.

One key to alleviating test anxiety is to reduce the importance of the results. Leaving out those tests whose scores really are important—minimum competency tests, the SAT, ACT and so on, what makes more routine scores so important to these children? Three influences are likely to have the most effect: parents, teachers and the students themselves.

Test anxiety can be especially severe in students who derive a great deal of self-esteem from schoolwork. They may be children who habitually do well in school, but who may not be particularly successful in other aspects of their lives. Academic success takes on a disproportionate importance.

Parents who put a great deal of emphasis on good grades can worsen this attitude. Apathetic parents are a problem too, of course; their children may come to consider school performance unimportant. But parental encouragement can go too far. When it makes the child so anxious to do well that the child does badly as a result, the encouragement is counterproductive.

Teachers, too, can contribute to test anxiety by overexhorting their students to success. The days and hours preceding the test are especially important in this regard. If teachers seem nervous about the test results, students will come to see the test the same way. Some teachers do, of course, have a great deal riding on the results—not only a sense of their own professional competence, but also the regard of their peers and supervisors. Communicating these concerns to the children, however indirectly, is likely to pull down performance.

Teachers' behavior in the testing room can also make a difference. Walking among the desks is inadvisable, unless it is absolutely necessary to prevent cheating. Teachers should never look over a child's shoulder. And, as mentioned earlier, unfamiliar faces in the testing room should be avoided whenever possible.

The main step to alleviating test anxiety, however, lies in helping students to see that the test results, while significant in certain aspects, are not of such importance that they should be cause for great concern.

In more serious cases, it may be necessary to involve parents. There is not much a teacher can do alone if the pressure is coming from

home. The discussion will require great tact, however. Particularly, it is well to avoid any suggestion that parents are pushing their children too hard.

A conversation with parents can usefully focus on how well the child is doing in school—and indeed, many children with severe test anxiety perform well in ordinary classroom work. The emphasis can be on helping the child to show his or her true capabilities in the tests. Parents will want to hear that the school's evaluation staff is familiar with the problem of test anxiety and take it into account when considering scores. Parents can also be told that for many of the tests given in school, the purpose is not individual judgment, but group results instead.

The most important contribution parents can make is to reassure the child that they love him or her as a whole person, regardless of how the child does on tests. Parents—and teachers as well—can also praise the student for accomplishments in other areas. If done consistently over a period of time, this can help the student not to rest his or her self-opinion disproportionately on school performance.

Teaching to the Test: Justifiable or Self-Defeating

Undoubtedly the most effective way to raise test scores—and also the way most likely to arouse controversy—is through instruction that closely tracks the test itself . . . "teaching to the test." It typically focuses on the narrowest possible skills that the test examines and includes classroom practice using items very similar to those on the test.

In some circumstances, teaching to the test is perhaps justifiable—for example, in preparing students for the SAT. There, some feel that what the test actually measures is not nearly so important as the score itself, at least to the individual student. But the practice can badly defeat the purposes of an evaluation.

Where an evaluation seeks to determine the effects of an ongoing educational program, a change in students' scores due to a few hours of special instruction will distort the results. Still, teachers with a particular interest in the evaluation's outcome may be tempted to "help" their students. This is why some evaluators do not make the tests available to teachers ahead of time. It is tactful to explain that some teachers "unconsciously" shift their instruction to conform to the test; and to avoid singling anyone out, the test must be denied to all.

If tests were better than they are—if tests could measure all the important outcomes of education, instead of just those particular skills that work well in the multiple choice—then teaching to the test would be no different from any other instruction. The fact that a special effort to teach to the test can raise scores is really an indictment of the tests.

Test-Taking Skills: Teaching How to Take a Test

Chapter 3 noted that every test measures a variety of skills besides those supposedly being evaluated, including reading ability, answer-sheet bookkeeping and calmness under stress. Here is a look at some of those skills more closely—skills applicable to multiple choice tests in general. They relate to the test itself, not to the subject matter of the test. Most are equally useful in tests of any subject.

Test-taking skills can be taught. Moreover, a number of studies show that instruction in test-taking skills can raise children's scores substantially. Some of the students who routinely do well on tests seem to have figured out these skills for themselves. There is no doubt that their classmates, too, can benefit from the same techniques.

Some educators oppose the teaching of test-taking skills. "If a kid knows the work," says one, "he shouldn't need all that."

However, first, the multiple choice test presents a highly artificial intellectual environment. The mere fact that learning these special skills helps to raise scores proves that. Nothing children learn in the ordinary curriculum helps them to cope with this environment. Leaving them adrift without the skills they need to cope seems unfair.

Second, some children have the skills anyway, which gives them an advantage. Teaching the same skills to others only equalizes that advantage.

Third, and most important in evaluation,

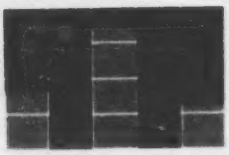
a lack of test-taking skills is part of the catalog of factors that prevent children from displaying on a test everything they know. Teaching the skills helps to remove this impediment and thus helps to make the evaluation more accurate.

What are these skills? For the most part, they focus on the special properties of the multiple choice test. For example, test authors have the responsibility of making one answer clearly right and the other three or four clearly wrong. Their efforts in doing so sometimes show through and help to indicate the correct answer. There is a strategy for guessing that can help to raise scores, even if the student is penalized for wrong answers.

Instruction in test-taking skills should not be saved until just before the test. It is more effective when spread out over the school year. One technique that will both help to improve test scores and reduce test anxiety: announce a test that will not count for anything. Explain that the answer sheets will be discarded without looking at them. Give the test—and in sight of the students, tear up the answer sheets and throw them away. Or, almost as good, insist that students do not put their names or any other identifying marks on their answer sheets. Either method can help students to become more familiar with the testing situation and to practice their test-taking skills with a minimum of anxiety. It also serves to put the testing experience in its place as just one of many rituals that children must undergo in school.

Chapter 5

Interpreting Test Scores: Looking Beyond the Numbers



By itself, a raw test score is not informative. Suppose one knows only that a certain student taking a certain test answered 60 items correctly out of 100. That says nothing. He or she may have done very well, very badly or anywhere in between. To learn anything useful from the score, one must be able to compare it with something else. Ask:

1. How does this student's score compare with other students' scores? A norm-referenced score answers this question. A norm-referenced score is not the test score itself—not the number of items answered correctly—but the child's standing among other students taking the test. Which other students?
2. How does the child's score compare with an educator's idea of satisfactory performance on the test? A criterion-referenced score answers this question. It reports whether the child achieved at expected levels, usually for several subtests within the instrument.
3. How does this child's score compare with what the same child achieved last time? Either norm-referenced or criterion-referenced scores are suitable for making this comparison.

There are many different ways to report results from the same test. As a result, test scores take a bewildering variety of forms. Each way of reporting a score serves some purposes better than others.

Criterion-referenced scores are usually self explanatory, though they cannot be fully informative without access to the actual test items. Knowing that a child achieved to a certain criterion tells little without a precise statement of the criterion. The most precise statements possible are the actual test items.

Norm-referenced scores always involve a comparison among students, and so the group used for comparison has a direct bearing on the results.

Test Standardization: Two Distinct Meanings

The term "standardization" has two distinct meanings in testing. First, it refers to students taking a test under uniform "standard" conditons—same instructions, same materials, same time available and so forth. Second, it refers to the process of developing norms or "standards" for use in reporting norm-referenced scores. For consistency, the word is used here only in the second sense.

The standardization process is one of the last before a test is released for general use. The publisher chooses a fairly large sample of students at the appropriate grade level. This group represents the national student population—the potential ultimate users of the test—in terms of age, sex, race, geographic location and so on. The students in the sample take

the test, and the publisher compiles their results.

When a student takes the test thereafter, a computer reads the answer sheet to find the number of correct answers. This figure is the "raw score." It is rarely reported at all on a norm-referenced test. Reported instead is an indication of how the child's performance compared with that of the standardization group.

Norm-Referenced Scores: Reporting Comparisons

A norm-referenced score compares the child taking the test to the standardization group. There are four common ways of reporting this comparison: percentiles, stanines, corrected scores and grade equivalents.

Percentiles are not, as some laypeople believe, the percentage of items a child answered correctly. **Percentiles** derive from a more complicated procedure. (See illustration.)

Suppose a particular child falls at the 60th percentile. That means he or she did better than 59 percent of the children in the standardization group and worse than 40 percent.

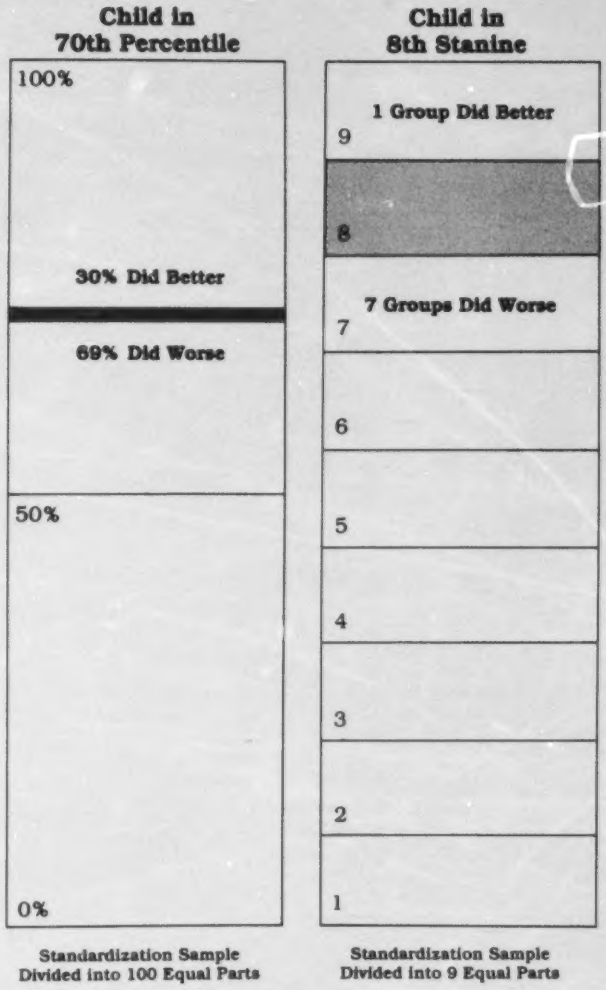
All tests show some error, however. If two children are one percentile apart, or even several percentiles apart, the difference is not ordinarily significant. In other words, the difference may well have arisen from chance error, not from any real difference in performance. Percentiles represent a finer degree of precision than a test can really yield. For that reason, many evaluators prefer to record scores in stanines instead.

The procedure for arriving at **stanine** scores is the same as for percentiles, except that the standardization sample is divided into nine equal groups instead of 100. A child in the eighth stanine, for example, is in the eighth of nine equal groups, counting from the bottom. Why nine? This particular breakdown approximates certain formulas that statisticians use. It also translates well into everyday language. The child in the eighth stanine, for example, is near the middle of the top third.

Corrected scores convey much the same information as percentiles or stanines, but they do it in different form. The publisher arbitrarily chooses a range of scores to use for reporting. The publisher must make two basic decisions: an average score, and how far the scores will spread out from the average.

Taking raw scores from the standardization group, the publisher finds their average and their spread then pegs the average corrected score—chosen arbitrarily—to the average raw score. Then, working out the math to obtain the desired spread among the corrected score, the publisher also pegs every other corrected score to a raw score. Often the publisher makes other adjustments so the cor-

Percentile and Stanine Scores



rected scores will follow a bell-shaped curve. The result is a table that translates raw scores into corrected scores.

There is one important difference between percentiles and corrected scores. Over the percentile scale, students are spread out evenly. There are just as many people in the 99th percentile as in the 50th percentile with occasional exceptions. But in most forms of corrected scores, students cluster heavily at the middle.

The final method of reporting norm-referenced scores is **grade equivalents.** They are easier to understand than the other three. The problem is that grade-equivalent scores are fundamentally ambiguous.

To compute grade equivalents, one needs an expanded standardization group. Suppose the test examines fifth-grade reading. The standardization group must now include not only fifth-graders, but also students from several grade levels above and below the fifth grade. All of these children take the same fifth-grade test. The publisher sorts the students by grade, and compiles a list of the average raw scores obtained by the students in each grade.

When a fifth-grade "customer" takes the test later on, the publisher looks up his or her raw score in the list. Suppose the student's raw score is 80. And suppose that sixth-graders in the standardization group, taking this fifth-grade test, averaged the same raw score of 80. This student's grade equivalent score, then, is six: the student is performing on the test as well as the average sixth-grader.

Because each grade represents a big jump in achievement, publishers usually report the grade equivalent in both years and months. As with percentiles, however, the difference of a few months may not be significant. Test error can cause score differences that do not represent real differences in achievement.

Grade equivalents are clear enough when the method for computing them is understood. The ambiguity arises when an evaluator says, "John is reading at a sixth-grade level." This can mean either of two things: (1) John reads fifth-grade material as well as a sixth-grader—the correct interpretation, or (2) John reads sixth-grade material—which is wrong; or at least, this test score does not say so.

Minimum competency goals are often expressed in grade equivalents, especially when formulated by state legislators who often lack much expertise in evaluation. A high school diploma might be conditioned, for example, on a student's ability to read and do math at an eighth-grade level. Presumably this means the student should handle eighth-grade material as well as the average eighth-grader. Probably most evaluators would interpret it that way. But the standard is meant for students in grades 11 and 12. Thus, it could also mean handling grade 11 material as well as the average eighth-grader, which is a different standard altogether. The ambiguity has caused some confusion in public discussions of minimum competency tests.

Grade equivalents are subject to another kind of problem. The theory assumes that students in higher grades will do better, on the same test, than students in lower grades. But this is not always the case—especially in the minimum competency context, where the standards are often set several grades below the students' current level. In tryouts of the Florida competency test, bright tenth-graders tested out in math several grades below their level. It caused a considerable furor at the time. But the answer turned out to be simple enough. The test included arithmetic with decimals, material taught long before grade 10. The students had forgotten how to do it.

Criterion-Referenced Scores: Each Skill is Tested

Criterion-referenced scores do not come in a few consistent forms, as non-referenced scores do. Each publisher has its own format. But criterion-referenced scores are usually self explanatory and relatively easy to understand. They also convey a good deal more information than norm-referenced scores do.

The report for an individual child lists a number of objectives. Each objective corresponds to a subtest within the instrument. Beside each objective is usually the following information, though not always in this order:

- the number of test items corresponding to the objective;
- the number of those items the child answered correctly;

- the number of items needed for "mastery" of the objective;
- whether the child achieved mastery.

There may be similar information that summarizes related clusters of objectives, and perhaps for the test as a whole. Some tests also purport to provide "diagnostic information"—indications of where the child needs special help. When available, this information is often given in abbreviations or code numbers. It requires the manual for interpretation.

Some criterion-referenced tests, especially those mandated by states, also give percentile information on the various objectives and on the test as a whole. But these percentiles are tricky. They cannot always be interpreted as if they came from a norm-referenced test. They may also be very confusing to parents.

On a norm-referenced test, the percentiles come from a representative standardization group—children in the grade level for which the test is intended. On a state criterion test, however, the percentiles usually come from all students taking that test that year. And the content of the test is typically set several grades below the grade level in which it is used. For state monitoring purposes, fifth-graders may be tested on second- or third-grade material. A large proportion of students, in consequence, will achieve mastery on a given objective. And a substantial number will answer every item for that objective correctly.

Parents find it upsetting to have their child bring home a test report with perfect scores and a mediocre percentile rating. An explanation like the one above may be necessary. And it will mollify many parents if the school offers to insert a written explanation in the child's file.

More important, if these tests play any part in program evaluation, the distortion in percentiles can lead to evaluation results that are seriously misleading.

Fifty-two percent of the responding districts in the AASA survey for this publication have competency tests mandated by their state or district. And of those, fully 85 percent use the competency test results in program evaluation. There are other arguments against using competency tests for this purpose. But it is sometimes risky to let percentiles from competency tests enter into program evaluations, because they can badly misrepresent students' performance. Caution is advised.

Averages and Errors: Considering Other Students

Most program evaluations make use of average scores. Fewer than four percent of responding districts in the AASA survey consider individual scores alone. All of the remainder, 96 percent, take averages into account. And over 60 percent of those look at nothing but averages.

Interpreting averages requires going a step beyond interpreting individual scores. Average usually is interpreted as representing a typical student—all students rolled into one, so to speak. But in practice it does not work out that way. Averaging is a mathematical process that can introduce distortions of its own into the data. Moreover, averaging can exaggerate certain errors in the individual scores.

Worse, some evaluators routinely take averages that are not permissible, mathematically—averages than cannot possibly have any meaning. The most common offense is to average percentiles. An "average percentile" makes no sense; it results from something like adding apples and oranges. Many districts report average percentiles nonetheless, and they often appear in local newspapers. There is in fact a correct way to arrive at a percentile figure that represents the entire student population. But simply averaging the percentiles is not the way to find it. See the accompanying illustration.

It is permissible to average measurements. It makes sense, for example, to compute the average height of the school's basketball team. It is also valid to average quantities—the average number of baskets scored per game, for example. But a percentile is neither a measurement nor a quantity. It is a third species of number, what mathematicians call a "rank." And averaging some kinds of ranks—including percentiles—is mathematically meaningless.

Averaging stanines is not valid either, for the same reasons. Nor does averaging grade equivalent scores make much sense, although

the reasoning there is a little different. It is, however, usually valid to average corrected scores, assuming that the publisher has set up the corrected scores properly.

But most of the tests used for program evaluation do not report in corrected scores. How, then, can we arrive at a percentile that meaningfully represents the average performance of the group? One method is to average the raw scores—the number of items that each student answered correctly—and find the percentile corresponding to that raw score. But there are two difficulties: many publishers do not report the individual raw scores at all, and even if they did, finding the average would entail a lot of dreary work with a calculator.

The best solution is to share the problem with the publisher, preferably before the test is administered. Ask for average raw scores, and their percentile equivalents, broken down however wanted—by programs, classrooms, schools or whatever factors are of interest. Depending on the publisher's computer set-up, this can either be very easy to accomplish (and hence inexpensive) or else quite difficult (and costly). The prices charged for this kind of service may become one factor in deciding which test to use.

Averaging Raw Scores: A Warning

An average that depends on raw scores may not be useful in comparisons from year to year. Publishers make occasional adjustments to their tests, and these can result in changes to the raw scores. Moreover, not all students taking the test at the same time will receive exactly the same test. For individual students, the publisher corrects for these differences when computing percentiles. But the correction is much more complicated when raw scores from different forms of the test are combined into a single average. This is not a serious problem when the test is given once; the results should be close enough. But in comparing scores from successive years, differences may appear that are due solely to changes in the tests. Talk to the publisher.

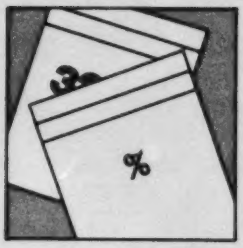
No information at all, however, is better than wrong information. Averaging the percentiles themselves is certain to give wrong information. As a rule, it will underestimate the performance of good programs—programs yielding scores better than the national average—and will overestimate the performance of poor programs. The distortion will be greatest when the scores as a whole are far removed from the national average in either direction, and when there is wide variation among individual students.

Even when done properly, averaging still introduces two other kinds of error. One is the "hole in the middle." Suppose a program is evaluated using a test that yields corrected scores like the SAT, so that averaging is permissible. Suppose further that a particular program is working quite well for some students, and very badly for others. Half the students score 300, and the other half score 600, with no one in the middle. The average score: 450. If only the average is considered, a uniformly mediocre program is seen. The average completely obscures the important question posed by this evaluation: Why the program acts so differently on different students.

Averages also overrepresent extreme performance, whether it is very good or very bad. A few scores that are much better or much worse than the others will "pull" the average up or down disproportionately. A better measure than the average, in such cases, is the "median." Sometimes reported in test score summaries, the median is merely the performance of the middlemost student—a student squarely at the 50th percentile. In many cases the average (the mean) and median will not be the same, though in educational testing they are rarely very far apart. (See Glossary.)

Individual Results: Another Personal Concern

All factors above suggest that an evaluation based on averages alone can be very misleading. Almost always, it is necessary to consider individual results as well. At the very least, looking at individual results—preferably in light of the actual test questions—can help to confirm that the averages mean what they say. While four percent of respondents to the AASA survey for this publication conduct program evaluations using individual results only, this is not practical for most districts. When



Two Ways of Reporting Scores

The Galveston (Texas) Independent School District reports its test data in both grade equivalent and percentile forms. Because the testing reported here took place in the eighth month of the first grade, the national grade equivalent average would be 1.8. The norms assume one month of progress for each month in school and one month over the three summer months—ten months per year in all.

TABLE 1
Iowa Test of Basic Skills
Mean Grade Equivalent Scores
Grade One

School	Number Tested	Word Analysis	Vocabulary	Reading	Math Total	Composite
Alamo	90	1.66	1.51	1.79	1.64	1.67
Bolivar	11	1.39	1.51	1.18	1.54	1.45
Burnet	99	1.74	1.42	1.58	1.50	1.58
Morgan	127	1.51	1.22	1.34	1.45	1.37
Parker	160	1.64	1.68	1.62	1.76	1.67
Rosenberg	107	1.93	1.77	1.84	1.82	1.82
San Jacinto	99	1.57	1.48	1.52	1.50	1.54
GISD	693	1.66	1.52	1.60	1.62	1.60

TABLE 2
Iowa Test of Basic Skills
National Percentile Rank of Average Pupil
Grade One

School	Number Tested	Word Analysis	Vocabulary	Reading	Math Total	Composite
Alamo	90	44	39	50	40	44
Bolivar	11	33	39	26	35	33
Burnet	99	47	35	42	33	39
Morgan	127	38	28	32	29	30
Parker	160	43	45	44	46	44
Rosenberg	107	55	49	53	51	52
San Jacinto	99	40	37	40	33	38
GISD	693	44	39	43	39	41

there is a lot of data, summarizing can be useful. But a look at the individual results can be a useful check on the summaries.

What to Look For: Inspecting Results

There is no one checklist for inspecting individual results that will work in every evaluation. There are too many variations in purpose, method of data collection, analysis and student body. The following is only a start. It stresses elements in the data that may point to an inappropriate test or to other major problems. The first three items below are grounds for serious doubts about the test results.

- **Bimodality**— more high and low scores than there are scores in the middle. Suggests either a test that is defective for the purpose, or a program that works especially well for some students and especially badly for others. Try to identify common characteristics among the students with high scores, and also among those with low scores.
- **Skew** —scores clustered near the high or low end of the scale and trailing off toward the other end. Suggests a test that is either too easy (high scores) or too hard (low scores) for the purpose. If the test has given good results in the past, skew may instead point to a program that is working exceptionally well or exceptionally badly.
- **Peaking (kurtosis)** —nearly all scores clustered in a very narrow range in the middle, with few high and low scores. Suggests an inappropriate test. The test is failing to distinguish among students. Scores may de-

pend on something other than the program under study.

- **High-end clumping** —good distribution of scores except for a clump or students high on the scale. May indicate cheating; identify the high-end students and check their classroom work. It may also point to a program that is exceptionally effective with certain students. Try to identify common characteristics among the high-end students.
- **Low-end clumping** —good distribution of scores except for a clump of students low on the scale. May suggest either a program that misses certain students, or particular trouble on the test—e.g., instructions not read properly in one classroom. Identify the students; check classroom work; identify their common characteristics.
- **Spot check** —Look up the test scores of several students whose work is known to be fairly consistent—some good, some indifferent and some poor. If any show unexpected performance on the test, either good or bad, investigate. Check to see if poor students succeeded consistently on any particular items, and if good students tended to miss the same items. Look at the items themselves to see if this is an expected result.
- **Extreme scores** —Investigate students at both extreme ends of the scale. See the discussion on "sensitivity at the extremes."

Sensitivity at the Extremes: One More Potential Problem

One additional kind of error has its greatest effect on very high and very low scores. Because extreme scores have a disproportionate effect on the averaging process, this kind of error can cause misleading averages as well. This is referred to as "sensitivity at the extremes."

Suppose a student answers one extra item correctly or gets one extra item wrong. What is the effect on her score? If the student falls anywhere near the middle of the standardization group, the one extra item will have very little effect. It may shift the score slightly but probably well within the test's margin of error.

If the student did very well on the remainder of the test, however—or did very poorly—the effect of that one extra item becomes much greater. It can change the score through several months of grade equivalent. A bright third-grader, for example, might be able to move from a fifth-grade equivalent up to sixth just by answering one more item correctly.

This effect, like several others examined, results from the bunching of students near the middle of the scale. In looking at individual results, it must be kept in mind that test error becomes very high for exceptionally good or bad students, much higher than for those whose performance is more typical.

Chapter 6

Minimum Competency Testing: A Special Method of Evaluation

In less than a decade, minimum competency testing has grown into a major influence on American education. Program evaluation is one of the routes through which it exercises that influence.

Three-quarters of the states have taken formal action on competency testing. By the early 1980s, 36 states had ongoing competency programs. Two other states had completed one-time studies. One state had action pending. Individual districts in many of the remaining 11 states had mandated their own programs.

Seventeen states flatly required students to pass the competency test before they could receive a diploma. Among these states are some of the most heavily populated, including California and New York. (Court action in Florida, one of the 17, enjoined the requirement until 1983.) An additional three states conditioned the diploma on competency scores less directly. Some states also made certain minimum test scores a condition for promotion from grade to grade.

The basic idea behind the tests in every state is the same: to ensure that the student has certain minimum skills. For high school graduation, the skills are typically set at an eight- or ninth-grade level. Some of the states examine "school skills," while others design their tests around everyday adult tasks. For example, one state might test reading comprehension from an essay; another would use want ads or appliance advertisements. One state examines rote arithmetic; another requires the same arithmetic in a problem involving wages or sales taxes.

The competency tests in every state examine reading skills. In all but a few they examine mathematics as well. Beyond these two subject areas, test content varies widely. Some of the more common topics include consumer awareness, job-finding skills and citizenship.

The Stated Purpose: Evaluating Students, Not Programs

The stated purpose of competency tests is not to evaluate programs, but to evaluate students in specified minimum skills. One state, Wisconsin, is considering enactment of a provision that would specifically forbid the use of competency scores for either teacher or program evaluation. A ban on program evaluation via competency scores appears nowhere else in the country. In the AASA survey for this publication 85 percent of the districts with competency programs use the scores for program evaluation. Some have expressed doubts about the wisdom of this practice, except for certain very narrow purposes.

The problems in using competency test scores for program evaluation arise from an

issue discussed in Chapter 3. To give meaningful results, an evaluation must use tests suited to the purposes of that particular evaluation. To assess the suitability of competency scores for program evaluation, one must look to the tests' original purposes.

The States' Reasons: Why Minimum Competency Testing?

The states that have acted on minimum competency testing express a variety of reasons for doing so. Among the reasons stated (some unofficially) are those below. Probably no one state has cited all the reasons here.

- To make the diploma more meaningful by having it represent at least certain minimum skills;
- To encourage students in acquiring th minimum skills;
- To help employers identify job candidates with minimum skills;
- To reduce the need for industrial entry-level training;
- To encourage individual teachers toward greater attention to the minimum skills;
- To encourage curriculum changes in the direction of basic skills;
- To identify students in need of remediation;
- To create a consistent data base for monitoring education statewide.

Minimum Competency Tests: What Districts Do

Fifty-two percent of the districts answering the survey for this publication have minimum competency tests mandated at either the state or district level. Of those, 61 percent use tests constructed by the state, and 30 percent, by the district. Some of those states and districts have help from publishers in test design. Another eight percent turn the whole job over to a publisher.

Competency tests are administered in an average of three different grades, but this average represents a wide range. Some districts test in every grade from kindergarten through twelve; some test only once, late in high school. The first test is given in grade 5, on the average; the last, in grade 10. Most districts indicate that a student who fails can continue to retake the test in later years.

The Test Results: Pressure to Pass

In those states and districts where minimum competency testing is the law, schools are under enormous pressure to have as many students pass as possible. A school building in New York state can even lose its state accreditation if too many students fail. In that light it is hardly surprising that many districts take the competency scores into account as part of their program evaluations.

As previously stated, a test can be satisfactory for program evaluation only if its objectives match those of the program being evaluated. Conversely, the choice of a particular test will favor programs with the same objectives as that test.

If competency tests indeed reflected a set of worthwhile educational goals, their use in program evaluation would make good sense. But few educators would be content with producing graduates who can merely pass the test. As a result, competency scores are limited as a form of program evaluation. Specifically:

1. Minimum competency tests examine a relatively narrow range of skills, most often just reading and math. Some programs are more extensive. While these skills are indispensable, they are only the foundation of an education. They represent far from an adequate education in themselves.
2. The required skills are established at a relatively undemanding level, typically three to four years below students' actual grades. Again, while this may represent an absolute minimum for a diploma, it hardly represents a satisfactory education.
3. Some states examine students in very specific applications of these skills—for example, filling out a driver's license application or computing overtime pay. The schools have traditionally taught more general skills that can apply to a wide variety of situations. Indeed, the hallmark of an educated person is the ability to solve new problems as they arise. Most competency tests do not recognize this ability at all, and so discourage the schools from

teaching in that direction. (Some tests do cover problem-solving skills.)

4. The tests are usually very easy for most students. Overuse of competency scores in program evaluation can lead to instruction that fails to challenge all but a few.

A Legitimate Use: Matching Tests to Objectives

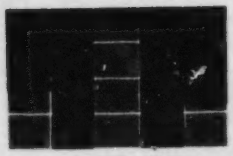
A test can yield valid evaluation results only if the test's objectives match those of the evaluation. Accordingly, a minimum competency test is useful for program evaluation purposes only if the program's objectives are the same as those in the test.

This condition is met for one narrow class of programs: those designed solely to help students pass the test. In most circumstances, this will be a remedial program offered to students who have already failed this test once, or who are in serious danger of failing it.

It is sensible to use competency scores as a measure of whether a remedial program is succeeding. The passing rate is the best indication of its success. But to the extent that the program has other, more ambitious goals, the test might be inappropriate as a device for evaluating it.

Chapter 7

Emerging Evaluation Methods: Emphasizing Students over Numbers



Of course, there are evaluation methods apart from those connected to a statistical analysis of test scores. Although a great many such methods are available, they are not used often—and even then, they often supplement test score analyses.

Circumstances often dictate the choice of method. When tests are already being administered for some other purpose and money is short, re-analyzing the same results for program evaluation may be the only option. Where evaluation personnel are scarce some of the qualitative methods may not be practical, since these approaches tend to require more staff time than test-based methods. Apart from the administrative factors, the popularity of tests may reflect the style of training most evaluators receive. The university programs leading to this profession typically stress testing methods and statistical analysis and give comparatively little attention to other methods.

Evaluating Without Tests: Quality Can Show

Deborah Meier, who has run the Central Park East School in New York for eight years, avoids using standardized tests for evaluation. She argues that the tests examine only a narrow range of abilities, that they are effective mostly for assessing groups, not individuals and that they measure extraneous phenomena—skills and abilities besides those supposedly being tested. These include, she says, various "sophistications" that are more common in some social groups than in others.

Meier says that test questions are designed to discriminate within a group, not to assess a child on his or her own merits. As many users and parents do not understand the mechanics of test construction, the tests are largely mystical. They distract, moreover, from meaningful educational goals.

Instead, Meier uses four techniques that examine student skills more directly than tests do. All four operate effectively with samples of children and need not involve entire groups.

1. Interviews using a standardized format, so that results may be coded and summarized.
2. Direct tests—e.g., having a child read a passage aloud or produce a writing sample. These results too can be coded. They enable the evaluator to explore specific strengths and weaknesses as a standardized test cannot.
3. Reviewing a standardized test with the child, item by item. The evaluator reads the questions aloud, and has the child explain his or her answers. Meier reports that this procedure often uncovers a higher level of skill than appears in the answer alone; often a child's "wrong" answers are

correct in some way when explained.
4. Collection of student material for inspection.

In addition, Meier's teachers write extensive reports on each child. They set objectives in writing and assess the child's progress against those objectives, in a manner somewhat similar to that used with individualized education plans (IEPs) for handicapped children under P.L. 94-142. Questionnaires also go out to parents.

This style of evaluation is time-consuming, Meier acknowledges, but all of the process directly benefits the child. All of the information cycles back; nothing is extraneous. No part of the procedure goes to "making others happy." The results, she stresses, are only suggestive; they yield a profile that gives a general sense of how the child is doing. There is no way to reduce these results to a small series of numbers, and she says it would be a false effort to try.

Methods Other than Tests: What Works for You?

The survey conducted for this publication confirms that evaluations using tests are the most popular. The next most common method is based upon reports from building principals. Only 54 percent of the responding districts report they usually employ this method. In-classroom observations are "usual" with only 50 percent; oral and written reports by teachers, 41 and 39 percent. The numbers drop off sharply from there.

On the surface it would seem reasonable to determine, first, what questions need to be answered through the evaluation. Then, in light of those questions, the most appropriate method should be chosen. In practice, however, the procedure often goes the other way: The method is chosen first, sometimes largely on the basis of cost or convenience. Only then, with the chosen method in mind, are the purposes of the evaluation determined. Some evaluators probably postpone deciding on the purposes even longer, until after the results are in. Whatever questions the data seem to answer most clearly are then deemed to have triggered the evaluation.

Simple Can Be Best: Ask Which Programs Work

The most common evaluation questions take the form of a simple comparison: Which program works better? Did students gain as much over the year as expected? These may in fact really be the important questions; or they may have come to seem important because the test-based methods can answer them with relative ease and at little expense.

Often, however, it is very helpful to go further. Qualitative methods can be useful for addressing a somewhat more difficult kind of question: **Why** did students do better in Program A than in Program B? **Why** were last year's achievement gains lower than the year before? Ordinarily test data provide no answers to these questions, although a detailed analysis of student answers, item by item, can help. Qualitative methods on the whole are more desirable.

Qualitative methods are also superior, even for simple comparisons, for certain classes of programs. Sometimes the program's goals do not lend themselves to a tidy list of objectives. And sometimes, even when the program can be expressed in a manageable number of objectives, those objectives may not show up well in a multiple choice test. In either case, imposing a test-based method may seriously misrepresent the program's results. The use of qualitative methods may then be essential.

Qualitative Methods: Doing Without Numbers

Qualitative evaluation methods, by definition, do not produce numerical results. The information they provide typically includes the following:

- Written reports by teachers, principals, classroom observers or students. These reports can be (1) in the writer's own words, (2) in answer to specific questions on a printed form or (3) in the form of completed checklists.
- Evaluators' notes on interviews with any of the above people. Again, the notes may be taken freely, on a form questionnaire or as a checklist.
- Examples of student work; or evaluators'

notes characterizing and summarizing student work.

- Tape recordings or transcripts of classroom interactions (often used for research program evaluations in mathematics); also, written summaries of these recordings or transcripts.
- Videotapes of classroom interactions; also, notes and summaries derived from the videotapes.
- Questionnaires returned by parents; notes and summaries from interviews and meetings with parents.

Gathering this kind of information usually presents no special difficulty. The problem is in what to do with it once it is collected. The material in its raw form—as a huge stack of forms and papers on the evaluator's desk—is of little help in decision-making. It is necessary to reduce the information to some kind of summary. Forms and checklists are very helpful for this purpose.

Many qualitative evaluations are ultimately reduced to numbers. The survey on evaluation practices, cited throughout this report, is one example. It requested no numbers except for data on budget and district size. But some of its results become interesting and useful only after their reduction to numbers. That survey represents a relatively simple case, however. Some evaluations, especially those depending heavily on checklists, present their final summary in a large table that shows the relative frequency of the various responses. Often graphs or charts are appropriate.

A qualitative evaluation will go far more smoothly if staff members decide ahead of time how they plan to summarize the information later. The task can be simplified by working backwards, from the form of the summary to the gathering of information.

Pilot Programs: Qualitative Evaluation Essential

Qualitative methods are especially useful in the evaluation of pilot programs—that is, programs being tried out on a small scale, with plans for later expansion hinging on an evaluation. Such programs often undergo intense evaluation, because large funding commitments can turn on the outcome.

Pilot programs nearly always look good in the evaluation. One important reason is the Hawthorne Effect, discussed previously. A second reason is that the program's objectives are often still in flux during the evaluation. Accordingly, whatever the evaluation shows to be the program's strong points become its main objectives, and the program is thereby deemed successful. Third, if the pilot program is breaking new ground, clear standards against which to compare it may be lacking. The promoters might be free to shop around for standards that will make the program look good by comparison. A fourth reason is that small programs have inherent advantages over big ones, especially if they interact with the community beyond the school. For example, a pilot program giving math students experience in local high-technology firms might be a resounding success with 20 students. But later on, with 2,000 students, there may not be enough suitable firms to go around. And finally, everyone wants to see the pilot succeed. It receives the kinds of attention and support that can make almost anything work.

With the pilot evaluation all but a foregone success, the important question becomes **why** it is successful--and whether that success will transfer to the program when it is run on a large scale later on. Here, test-based methods are of little help. Qualitative techniques are nearly always necessary. It takes an insightful evaluator to ask the right questions and to construe the answers properly. But there is usually no other way to ensure that a small success does not turn into a large disappointment.

Checklists and Forms: Trying Them Out Helps

Whenever forms or checklists of any kind are used, it is usually advisable to "pilot" them before the evaluation. Simply have a few people try out the forms in situations similar to the planned use. For example, if the form is for classroom observers and bears on, say, student-teacher interactions, a small number of observers can test the form in actual classrooms ahead of the main evaluation. This not only ensures that the form is convenient to use—a most important consideration—but that

An Evaluator's Checklist

This checklist is an excerpt from a booklet developed by the Willow Run Community Schools in Ypsilanti, Michigan. Additional parts of the booklet, all in similar format, look to school organization, students, media and materials center, administrative practices and procedures, school maintenance and food service.

						CONDITIONS COMMENDABLE
						CONDITIONS ABOVE AVERAGE
						CONDITIONS ACCEPTABLE BUT IMPROVEMENTS NEEDED
						CONDITIONS NOT ACCEPTABLE
						SEE COMMENTS— LAST PAGE
						NO COMMENTS OR UNOBSERVED
					INSTRUCTIONAL PROGRAM Program is responsive to district goals.	
				Program is designed to meet various abilities, talents and interests of the students.		
			Remedial programs are provided where needed.			
		Accelerated programs are provided where needed (reading, math, gifted, etc.).				
	Testing program results are used for improvement of instruction.					
The school has a positive learning environment.						

the form allows reporting the events and characteristics that the observers see as most useful in the evaluation.

Where several observers are involved, it is also advisable to run a check on their consistency. This check can often occur in conjunction with pilot-testing the forms. In the example above, several observers can monitor the same classroom, and can later compare their forms. Serious discrepancies might suggest the need for clearer criteria, different forms, or better training of observers. Observers should be encouraged to discuss their different impressions. This can help to make their later observations more consistent.

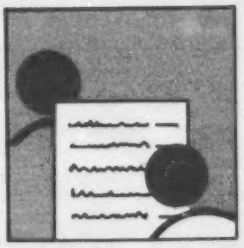
Classroom Videotape: One Way to Evaluate

Children will forget a television camera even more quickly than a live observer, after one or two questions about it. The camera may also be less intrusive for the teacher. Videotape has the further advantage of allowing observers to check each other's work, which promotes consistency. The technique is not expensive for a district that already owns the equipment, but it may be prohibitive for others.

Teacher/Principal Reports: Both Popular Methods

These are among the most popular qualitative methods, according to the AASA survey. A few suggestions:

- **Use forms.** If more than a few teachers and principals are involved, the results will almost certainly be unmanageable without forms.
- **Keep the forms short.** Resist the temptation to add questions. A few short forms will probably give a better response than one long one.
- **Make the questions very clear,** so respondents know precisely what is being asked.
- **Consider the use of rating scales** where appropriate. (For example, "check one: strongly dislike, dislike, don't care, like, strongly like.") Rating scales are fast to complete and, if properly designed, they give a lot of information. More than seven choices are usually too many, and five are often enough. Examine the choices offered to make sure the wordings are evenly balanced.
- **Check all questions** to ensure they do not prompt one answer over another. Most people filling out forms tend to give the answers they think the researcher wants.



A Sample Interview Form

1. What is the range of topics to be explored during your course?
2. As a result of your course, what skills, attitudes or knowledge do the students develop?
3. Do these topics, skills and attitudes seem to you to be appropriate for the age student with whom you work? If you think some other areas should be stressed, what are those skills or topics?
4. Do you integrate your teaching of social studies with any other disciplines?
5. Are you in a self-contained or team teaching situation?

Source: Excerpted from a form supplied by Ted Nicholson, Wassau, Wisconsin

Personal Interviews: Allowing Individual Input

In addition to gathering written information, interviews serve the useful function of allowing people the satisfaction of making additional input. In the evaluation context, it is often useful to plan the interview in two parts. The first part is tightly structured; the interviewer follows a questionnaire or checklist. The second part is more open-ended, giving the interviewee more control and allowing her to raise her own points of concern. When students are interviewed, they may need con-

Surveying Students and Parents

Pupil Questionnaire

Pupils were asked to respond to various items concerning the school climate and activities. Responses were received from approximately 60 fourth, 70 fifth and 50 sixth graders. Among these pupils, approximately 20 were Native American, 30 were black and 80 were white. A composite of the results compiled by grade and by race follows.

1. Did you enjoy being at Roosevelt this year?

				Percent of Responses		
Response	**Gr. 4**	**Gr. 5**	**Gr. 6**	**Native American**	**Black**	**White**
A Lot	74%	43%	40%	62%	55%	53%
A Little	26%	47%	52%	38%	35%	45%

2. Did you make any new friends this year?

Percent of Yes Responses					
Gr. 4	**Gr. 5**	**Gr. 6**	**Native American**	**Black**	**White**
98%	100%	100%	100%	100%	99%

3. Do you want to be at Roosevelt next year?

Percent of Yes Responses				
Gr. 4	**Gr. 5**	**Native American**	**Black**	**White**
90%	46%	50%	67%	82%

4. What do you like best about Roosevelt?
   - The most frequent responses were swimming, teachers and other physical education activities.
5. What is your favorite class?
   - The most frequent responses were homeroom, gym, art, library, science and music/piano.
6. If you could, how would you make Roosevelt better?
   - The most frequent responses concerned fighting and litter in the building and on the grounds.

Parent Questionnaire

To determine parent attitudes about the program, a questionnaire was sent home in the spring. The responses from 115 parents follow.

1. Are you pleased with the program at Roosevelt?
   - The responses were 54% "very pleased," 43% "generally pleased" and 3% "not pleased."
2. If you could choose any elementary school at all for your child to attend next year, what would it be?
   - The responses were 80% Roosevelt, 14% another Tulsa public school and 6% out-of-district or private school.
3. What do you like best about Roosevelt?
   - The most frequent responses concerned the quality of the staff and the quality of the curriculum.
4. If you could, how would you improve the program at Roosevelt?
   - The most frequent response concerned increased parent involvement.

Source: Excerpted from a report issued by Paul I. McCloud and colleagues in Tulsa, Oklahoma

siderable encouragement in the second stage. But they may also raise issues that would have been lost otherwise.

Assessing Student Work: Seldom Used, but Effective

Assessment of student work is used relatively seldom, according to our survey, but it can provide a very direct reading of the program's effectiveness. It is especially informative to take samples of student work at different times throughout the year and examine them together. To make the student samples easier to compare, evaluators can suggest specific assignments to teachers for these occasions.

Involving Parents: A Positive Payoff

Some school districts involve parents in the evaluation process, either through meetings or via questionnaires. The political benefits of giving parents a voice can be a valuable side-effect. And many parents are quite knowledgeable about educational practices and trends. Some school personnel, however, show a tendency to "talk down" to parents, or to use educational jargon that parents do not necessarily understand. Both tendencies must be resisted, or the contact can be counterproductive. The best posture is one of equals to equals, two adults concerned with the welfare of the same children. Both parties know that the ultimate decisions are the school's, but parents will appreciate a respectful attitude and clear explanations, nonetheless.

Respect for Children. Perhaps Most Important

Indeed, respect may well be the single most important factor in all of evaluation. First must come an evaluator's respect for children. Without that, the tables and charts can give at best a poor picture of the children they purport to represent. It takes respect and sensitivity to make an evaluation that yields good data in the first place and even more to look beyond that data to the children whom the data stand for.

This is only the beginning of the chain, however. The evaluator too needs respect from teachers, principals, curriculum staff—all the way up to members of the school board, beyond to the parents who elect them and closing the cycle, back to those parents' children. More than almost any other endeavor, education is a close relationship of people at many hierarchial levels, each contributing his or her own specific skills and concerns.

In this complex relationship, the evaluator occupies a key pivotal point. The evaluator, in a word, cycles information from the ultimate goals of the process—children's learning—back to the making of decisions on how best to bring about that learning.

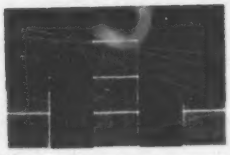
Engineers call this kind of function a "feedback loop," and they treat it with the greatest respect. They know that even minor disturbances in the feedback can send a large, smoothly functioning system into wildly erratic behavior, sometimes into self-destruction. Physicians too have learned the devastating effects of small damage to the body's feedback mechanisms. Social scientists are beginning to understand how feedback operates in systems made up of people. There, too, the feedback role can have a disproportionate effect on the success or failure of the entire enterprise.

The Evaluator's Role: Understanding and Sensitivity

In the school context, one conclusion is already very clear. The evaluator holds a special place in the ongoing search for educational quality. Evaluation provides the "leverage" in education. Evaluation is where a small push can have the greatest effect, where a single hand can most easily turn the massive educational machine. The push must be properly applied, however, or the whole system can go awry. Conversely, the right push can improve its performance greatly and with little effort. That makes evaluation a critical responsibility. It is fortunate indeed that so many evaluators understand the responsibility and are able to meet it with the intelligence, the sensitivity and the respect it demands.

Chapter 8

Evaluation to Improvement: A Never-Ending Process



Renewal is basic to the success of any institution. The person or organization that stands still can expect to fall behind. Why? Because society is changing. The pluralistic nature of most communities is evolving.

In some communities, demographic changes mean increasing numbers of older people, many without students in our schools; increasing numbers of Hispanic students . . . the list goes on. The very economic base of our nation is evolving, some say, from an industrial to an informational society.

Industry in the United States is currently undergoing a major transformation which may impact our economy with as much force as the industrial revolution initially had on the agricultural economy, explains Roy Forbes, director of Assessment and Evaluation Division, Education Commission of the States in Denver, Colorado.

"Among the factors responsible for this transition are the technological innovations in electronics and computer systems, a shift away from energy-intensive production and increased competition for world markets," Forbes says.

The Bureau of Labor Statistics predicts that the computer industry will lead all other industries in terms of product output increase throughout the next decade. As new uses and markets for computer technology become apparent, employment will grow most substantially at the upper-skill levels and least at the lower-skill levels.

During the period of 1949-1965, 8,000 types of jobs disappeared from America's labor market, largely due to automation. At the same time, more than 6,000 new types of jobs appeared. It is predicted that by 1990, more than half of the labor force will be employed in technical and white-collar fields.

Because an upgrading of labor force skills is imminent, many legislators and educators are rightfully concerned about preparing today's youth, as the future workers of our country, for entrance into the employment market. Industries can assume some of the responsibility for retraining experienced workers to move laterally within their firms. They will not, however, assume full responsibility for all unemployability problems, particularly for workers whose basic educational backgrounds are insufficient.

Forbes predicts that a typical job description in 1990 will likely include many tasks such as these:

- Monitor work flow of production systems.
- Prepare input for and execute utility programs.
- Provide interface with other departments about possible enhancements and determine computer program errors.
- Provide technical liaison with users.

- Evaluate and maintain new software/hardware.

This list is by no means exhaustive and not intended to represent the job level of every future employee. Tasks such as these, however, will become more widespread across all industries. Obviously the more technical the job, the more on-the-job training that will be required, although educational systems can provide essential preparation for the workplace.

Forbes adds that the "basics" of tomorrow will be the skills that today are considered to be of a higher-level. These skills include:

- Evaluation and analysis skills
- Critical thinking
- Problem-solving strategies, including mathematical problem-solving
- Organization and reference skills
- Synthesis
- Application
- Creativity
- Decision-making given incomplete information
- Communication skills

How are today's students currently performing on some of these higher-level skills? Forbes questions. How will our schools assure that educational programs are available to accommodate these changing needs?

For the most part, we have discussed the evaluation of existing programs. But there is another step. From time to time, it is a good idea to take a look at the total educational program to assure that it makes available learning opportunities that will help meet the real needs of students today and in the future.

When the American Association of School Administrators undertook a process of long range planning, its leadership identified seven **external factors** that would impact on the schools and the school administration profession over the next several years. Those external factors are: the labor market and economic conditions, social and demographic context of school systems; expectations and the role of the educational system, human and technical resources, financial resources and costs, energy, and organizational and legal constraints.

A school system might ask, "What effect will these factors have on our community?" "What effect should they have on our educational program?" "What should be added to or deleted from existing curriculum?" "What should we be offering that we are not?"

Schools are very much a part of society. They influence it and are influenced by it.

No school system can survive and prosper if disconnected from the community. School systems are funded, accredited, regulated, scrutinized, pushed, cajoled, studied and judged by others. Parents want to be involved in their children's education. Nonparent taxpayers want some assurance that their tax dollars are being well-spent. State and federal education agencies require the involvement of special populations or the community at large as a condition for receiving grant monies. All these factors impose special challenges—and offer specific opportunities—to evaluating educational programs. In general, these types of community evaluations fall into two categories: those *imposed* on a school from outside and those *invited* by a school.

Some federal programs have mandated formal community involvement in educational programs through such methods as advisory councils. State accrediting institutions, too, frequently require that a school district show evidence of "community input." School districts that have used this model successfully offer several suggestions:

1. **Make sure the group understands its function.** If possible, have the school board or appropriate administrators meet directly with the group to present them with their charge. Above all, ensure that the group understands that it is not to become a second school board.

2. **Have the group adopt planning premises.** If their function is to help the school district with long-range planning, have them formulate some thoughts on what changes they think will occur in society at large during the next five years.

3. **Have the group identify educational goals that relate to their planning premise.** One school administrator was pleasantly surprised that an advisory group that had

started out by expressing considerable concern over teaching "the basics" eventually identified only one goal that specifically dealt with traditional academics. Because of the thinking the group had done about the needs of living in the future, many of the other goals dealt with sometimes less tangibles such as learning to cope with change, career awareness, and so on.

4. **Develop objectives to meet the goals.** Depending on the mandate of the group, this may be a staff function or it may be shared by staff and community members. It is the responsibility of the staff, in any case, to help the group identify objectives that are in some way measurable.

5. **Use the objectives to evaluate the program.** Even with groups that are charged *only* with an evaluation function, the first four steps are critical to ensuring that the evaluation will be usable.

Mandated Groups: Using the Results

Probably every school district has at least once gone through a process of soliciting community involvement, has drafted some elegant-sounding objectives, has written a finely-crafted report, has sent the report off to the appropriate agency and then has filed the report away on a bookshelf where it gathered dust. But according to many administrators, the real challenge facing educators begins at this point—making the report a springboard for action.

The key, of course, is educational leadership. The commitment to making the process more than a dry exercise begins at the very top—with the superintendent in a school system, with the principal in a building. In part, the issue may be resolved by changing the focus of the question from "What do we need to do to satisfy this state or federal reporting requirement" to "What do we need to find out for ourselves, and how can we develop something that will satisfy these requirements as a byproduct of that inquiry."

Accreditation Efforts: Voluntary Self-Evaluation

Although membership in a regional accrediting agency is voluntary, going through the accreditation process is in a sense another type of evaluation imposed from outside. The comprehensive evaluation must be completed every seven years. Few schools that are not members of such an accrediting association ever undertake the kind of rigorous self-study necessary for accreditation.

Some schools have been so enthusiastic about the accreditation process that they have voluntarily invited members of the visiting team back a year later to help them evaluate their progress. Other faculties have regarded the entire process as an imposition, and have finished with it as quickly as possible. The difference seems to come from the leadership imposed by the principal of the school. If accreditation is viewed primarily as something that outsiders do *to* the school (a misinterpretation, according to accrediting association officials), staff members are likely to view it warily, if not with outright alarm. If, on the other hand, the accreditation process is seen largely as a self-evaluation, and the visiting team viewed principally as additional resources to help validate the study results, the outcome is far more likely to be positive and useful.

Accrediting Agencies: One Example

Six regional accrediting associations operate nationwide. The North Central Association, which includes 19 states, is the largest of the six. These voluntary associations are unique accrediting institutions, according to Dr. Cile Chavez, director, Colorado State Committee, North Central Association, because they use standards established and reviewed by members. The standards cover everything from how many credit hours are required for teachers wishing to teach certain subjects to the appropriate ratio of students to guidance counselors. While these standards do not in themselves guarantee educational quality, the members believe they provide the conditions that best assure the development of educational quality, Dr. Chavez said.

Community Surveys

In November 1981 Mesa Public Schools in Arizona sponsored a district-wide community survey to assess citizens' attitudes, which provided data to be used for planning the future direction of the district.

Since the survey was distributed throughout the Mesa school district on a random basis, not every household received a questionnaire. In fact, approximately nine percent of all households were given a questionnaire. Those persons who did not receive a questionnaire were given the opportunity to have one mailed to them upon request. In addition some surveys were distributed to groups such as civic clubs, etc.

*1981 Community Survey by Mesa Public Schools, Department of Planning, Staff Development and Special Projects, Dr. Bill Raymond, director. Dr. George Smith, Superintendent.

How Would You Grade Our Schools?

The question:

Students are often given the grades A, B, C, D, or FAIL to denote the quality of their work. Suppose the public schools in Mesa were graded in the same way. What grade would you give the public schools here—A, B, C, D, or FAIL?

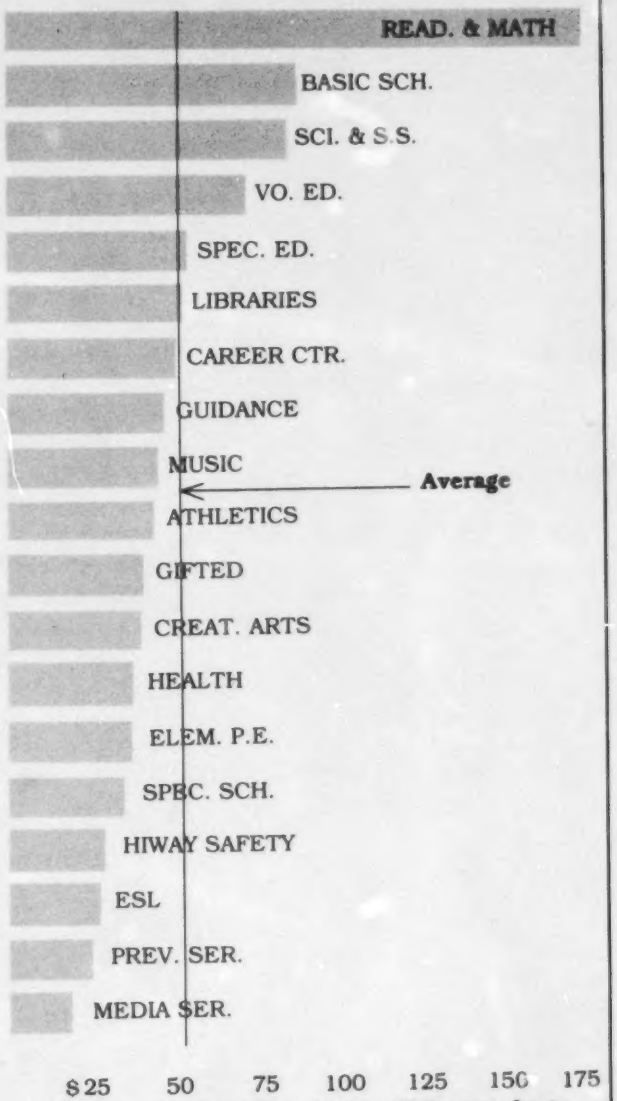
	A %	B %	C %	D %	Fail %
TOTAL 1981 Mesa Total	22.1	50.0	21.1	5.3	1.5

It is interesting to note that in the 1981 Gallup Poll only 36 percent of the respondents, in response to the same question, gave the public schools nationally either an A or B rating. The Gallup Poll provided this same information for the western states. ...espondents in the west rated their schools lower than for the entire country. In the western states 28 percent of the respondents gave the public schools either an A or B rating. In Mesa over 72 percent of the respondents gave the schools either an A or B.

	A %	B %	C %	D %	Fail %
SEX					
Men	21.5	47.4	23.1	6.2	1.8
Women	23.1	50.5	20.6	4.9	.9
RACE					
Black	0	0	50.0	50.0	0
Anglo	22.2	49.9	21.8	5.0	1.1
Mexican-American	22.2	50.0	22.2	5.6	0
American Indian	7.7	46.2	23.1	7.7	15.3
Asian	50.0	37.5	0	12.5	0
AGE					
under 25 years	14.3	55.3	28.6	1.8	0
25-45 years	21.8	51.7	20.3	4.8	1.4
45-60 years	24.7	45.1	20.9	7.6	1.7
over 60 years	22.9	46.9	22.8	5.1	2.3
RESIDENCE					
Residence Status					
Full-time	22.6	50.1	20.6	5.3	1.4
Part-time	11.5	48.1	32.7	5.8	1.9

	A %	B %	C %	D %	Fail %
Length Of Residence					
Less than one year	28.6	48.0	19.5	2.6	1.3
One to three years	17.9	53.1	24.6	2.9	1.5
Three to ten years	23.4	52.3	17.2	6.2	.9
Over ten years	22.8	45.8	22.8	6.5	2.1
ASSOCIATION WITH THE DISTRICT					
School Employee					
Yes	37.5	48.8	10.0	3.7	0
No	21.0	49.6	22.6	5.5	1.3
As a student attended MPS					
Yes	17.9	49.2	25.1	6.7	1.1
No	23.1	49.5	21.0	5.1	1.3
Have children currently attending MPS					
Yes	21.9	54.1	18.1	5.3	.6
No	22.3	45.5	24.4	5.4	2.4
Have had children who previously attended MPS					
Yes	21.9	51.6	19.4	5.7	1.4
No	23.0	46.8	23.8	4.6	1.8
Have children attending private schools					
Yes	12.7	36.4	32.7	9.1	9.1
No	22.7	50.2	20.8	5.1	1.2

Comparison Of Perceived Importance For Various School Programs



Average Amount Allocated By All Respondents

Accreditation is a three-step process. The first, and most involved, is the **self-study.** During a one-year period, students, faculty, and parents take a comprehensive look at the entire school program. They first explore two areas:

School and Community—What is the school's community? What does the community expect of the schools? How are the schools meeting those expectations?

Philosophy and Objectives—What is the school's philosophy? What objectives has it identified for meeting that philosophy? Several less extensive studies follow: Educational Program, Student Activities, Student Services, Learning Services, Facilities, Staff and Administration, Communication, and Emerging and Unique Programs. Schools are encouraged to develop alternative self-study formats, and the association helps them find resources to assist them in their self-study.

The second phase of the accreditation process is the **on-site evaluation by a visiting team** of educators from other districts. These professional colleagues both validate the self-study and suggest additional ideas for fulfilling a school's basic goals and objectives.

The final phase of the accreditation process is the **followup.** Within a year after the team visit, schools must submit a plan outlining how they plan to implement the recommendations that have come out of the study.

Community Needs: How Schools Respond

In addition to the evaluations that are imposed on schools from outside, most districts also invite community members to evaluate how effective the school district is in meeting the needs of society. These evaluations may involve some or all of the community; they may be formal or informal; they may cover a specific subject or may be all-encompassing; and their results may be released publicly or may be for internal use only.

Surveys. School administrators generally feel comfortable using a questionnaire to help identify educational problems and solutions. In fact, a well-designed survey can be a very useful evaluation instrument. A poorly-designed survey, however, may be worse than nothing, because it gives the illusion of providing "hard data." Sally Keeler, who has conducted several surveys for the Montgomery County Public Schools in Maryland stresses the importance of ensuring an adequate response from a random sample of the community. "Otherwise, you may have some useful information, but you certainly don't have survey results," she cautions.

One other administrator, who has designed surveys for a large school district, notes that it is important to remember that "in all questionnaires, there are a number of questions we could ask, but we don't because we really don't want to know the answers."

Many formal studies now include several of the questions from the Gallup Poll on education. Because the Gallup organization has a data base of more than ten years of national studies, these questions provide districts a chance to measure their own community against national norms.

Surveys should offer respondents the opportunity to define what they mean by such catch phrases as "the basics." Keeler noted that many community members who said the schools should be doing more to teach "the basics" were actually fairly pleased with the way the school system was teaching basic academic subjects. They expanded the definition, however, to include more traditional teaching methods, more emphasis on discipline, and a general nostalgia for the way things used to be. For examples of community surveys, see the illustration accompanying this chapter.

Informal surveys. Formal surveys, because of the need to ensure randomness and adequate response, are frequently rather expensive. Many districts, however, have found less formal methods to solicit information from the community.

- Value questions, such as "What three areas do you think should be given greatest emphasis during the coming year?" may be included in a formal or informal attitude survey conducted by the school system. These attitude surveys are often conducted by the school district public relations office and sometimes by local schools.
- Parent-teacher groups may be surveyed to

explore their attitudes toward what the school is doing, what it should be doing, and how well it is doing. Sometimes these questions are included as a "school report card" sent home along with the child's report card. Sometimes the questions form the basis of discussion groups that meet during a regularly-scheduled meeting of the parent-teacher group or advisory committee.

- Groups that may have been formed initially to satisfy state or federal requirements may be asked to continue providing regular information to schools.
- Everyone in the school community—faculty, students, parents, administrators, and community members—may be asked to list several objectives for the school. Then the groups may be asked to list which objectives should be most important; then which objectives seem to be given greatest importance. By examining the disparities within and between groups, administrators can find useful data for further planning.

Task Forces. When a school district decides to evaluate a specific curriculum area intensively, it will frequently convene a curriculum task force. These task forces traditionally include faculty members, parents, and administrators. They may include college or university professors. The Jefferson County, Colorado, public schools have also invited professional scientists and "consumers of science" to serve on their district-wide science curriculum task force.

School Improvement Councils: Making Schools Even Better

Some schools have appointed school improvement process councils to study community needs, become knowledgable about the curriculum and make evaluative comments to the administration and board for consideration in making schools even better. In some cases, school improvement process councils, or accountability committees, are mandated. These committees often include: administrators, parents, teachers, non-parent taxpayers and others. Some school systems have developed local school improvement councils. At these schools, the principal pulls together the staff and community representatives to regularly explore the educational program. The result of this activity is sometimes a local school improvement focus for the year. These councils can also help connect formal and informal evaluation to actual school improvement and, at the same time, build support for that improvement.

Effective Communications: Essential to Evaluation

The results of all these evaluation instruments—whether imposed or invited, formal or informal, frequently reveal that most citizens do not know very much about what the schools are doing. And that can cause real problems for schools. In fact, George Gallup traces part of the decline in public confidence in education to a lack of communication between school districts and taxpayers. Therefore, an effective communications program is an essential component of any evaluation of educational programs.

First, communications programs are essential because they must give the community information on which they can base their opinions about the school's programs. Gallup's research has led to a firm belief that a school district that relies primarily on the mass media—rather than its own communications program—to disseminate information about the schools will by and large have a more critical community, primarily because of the nature of news. The dramatic event or statistics are often considered more newsworthy than stories devoted to describing steady progress.

Once the evaluation has been completed, the community surveyed, the advisory council convened, and the data gathered, it is up to the school administrator to decide what to do with the information. How can what has been learned be channeled back into the process of improving educational programs?

George Redfern, former associate director of AASA and a consultant on educational issues, suggests that the first step is to convene a meeting of all those who will be charged with carrying out any proposed changes. The departments of instruction, personnel, finance, communication, as well as building principals should certainly be included. Redfern notes

that making the group inclusive rather than exclusive can lead to greater acceptance of the plan developed.

The group should address several questions.

1. **Do we want to change?** Several administrators noted that finally all the instruments in the world must pass one final test: do the results seem to make good sense? (This, of course, is why administration is both a science *and* an art.) The decision-makers may simply decide that the community does not understand why the school has adopted certain programs or policies. In that case, the decision may well be to continue with the program for the time being and to increase communications efforts to let the community learn more about what is being done.
2. **Can we change now?** Particularly in the middle of the academic year, it may be difficult to reallocate staff and budget to make any significant changes in programs.
3. **If not, when can we change?** Because school district budgets are so heavily weighted with personnel costs, planning for any major program changes may also include personnel changes. The sooner these plans can be begun, the better.
4. **Whatever our decision, how can we best communicate our intentions to respond to the community?** At this point, the evaluation process is really about to begin again. For as the district conveys its plans for program changes, it will be establishing community expectations by which those programs will eventually be judged. And that evaluation process will again lead to new programs.

Ultimately, evaluation of educational programs today should result in a better education for students tomorrow.

Sample Policies

Sample Policy* Evaluation of Instructional Programs and Outcomes

The Board believes that accountability for pupil performance and progress is a shared responsibility of teachers, administrators, parents, and the Board. Individual pupil progress and the instructional efforts of the schools shall be evaluated systematically.

The purposes of the evaluation process shall be:

1. To monitor the progress of individual students and periodically report that progress to the student and the student's parents.
2. To identify strengths and weaknesses of existing instructional programs.
3. To provide data for decision-making regarding additions to, modifications of, or deletions from the existing instructional programs.
4. To report to the public the relationship between the stated educational goals of the District and the achievements of the District.

The professional staff shall provide continuous evaluation of the educational program and instructional processes for the Board of Education. Acceptable data sources shall include, but not be limited to: standardized testing, criterion testing, teacher observation, attitude surveys, teacher judgment, and student self-evaluation.

District Size: K-12; 80,000 students

*Sample policy from Jefferson County School District R-1, Lakewood, Colorado, furnished by the Educational Policies Service of the National School Boards Association

Sample Policy* Evaluation of Instructional Programs

The Board has ultimate responsibility for the educational programs of the district's public schools. In order to appraise the effectiveness of the educational programs and instructional processes, individual pupil progress and the instructional programs and efforts of the schools shall be evaluated systematically.

The purposes of evaluation of instruction shall be:

- to determine educational needs and provide information for planning.
- to indicate instructional strengths and weaknesses.
- to check on the suitability of programs in terms of community requirements.
- to show the relationship between achievement and the system's stated goals.
- to provide data for public information.

The Board shall rely on the superintendent, administrators, and the teaching staff to provide continuous evaluation of the educational programs and instructional processes. Outside agencies may also be asked to evaluate the district's instructional program.

Evaluation shall be based on standardization instruments (such as testing programs) and staff appraisal of pupil achievement, in terms of individual, local, and national norms, and on such other criteria as may be developed.

District Size: K-12; 4,500 students

*Sample policy from Forest Hills Public Schools, Grand Rapids, Michigan, furnished by the Educational Policies Service of the National School Boards Association.

Glossary

Analysis: see Data analysis.

Bias: see Test bias.

Bimodality: describes a score distribution in which there are more high and low scores than there are scores in the middle; a distribution that shows two peaks.

Coding: the process of summarizing results in numerical form.

Concurrent validity: the extent to which the results of a test agree with the results of other measures administered at about the same time.

Construct validity: the extent to which the results from a test support the inference of a hypothetical characteristic thought to vary among students; e.g., intelligence.

Constructed-response test: a test in which the child must supply his or her own answers. Distinguished from multiple-choice test.

Content validity: the extent to which a test's objectives match those of the underlying curriculum.

Control group: children who, though not taking part in the program being evaluated, are tested similarly to those who are taking part for purposes of comparison. Distinguished from experimental group.

Corrected scores: an arbitrary scale for reporting test scores, usually norm referenced and adjusted to fit the normal distribution. Both IQ and SAT are examples of corrected scores. Distinguished from raw scores.

Criterion-referenced test: a test that reports scores in comparison with fixed standards of performance. Distinguished from norm-referenced test.

Criterion selection: the separation of children into experimental and control groups according to some criterion, such as previous test scores.

Curricular validity: the extent to which the subject matter of a test appears in the underlying curriculum; compare instructional validity.

Curriculum (as used in this report): the content, sequence, pedagogical style and materials for a course of study; includes all aspects of the instructional process under educators' control.

Data analysis: the process of converting raw data from test scores or other sources into a form suitable for decision-making. Often this involves numerical computation of averages, ranges and statistical significance.

Design: see Evaluation design.

Distractors: the wrong answers on a multiple-choice test.

Distribution: a graph of test scores that shows the proportion of children who achieved each score.

Equivalent forms: a method of assessing test reliability that compares scores from two different forms of the test thought to be equivalent.

Evaluation design: the specification of instruments or other data-gathering devices for use in evaluation, methods of their use and data analysis needed. A sound design ordinarily depends on a clear sense of the purpose the evaluation is to serve.

Evaluation policy: a written statement of a jurisdiction's intent as to evaluation, typically including broad purposes, subject matter, time intervals, audiences, decisions to be supported, reporting requirements and limitations, applicable standards and personnel responsible.

Experimental group: children who take part in the program being evaluated and are tested for the evaluation. Distinguished from control group.

Face validity: the extent to which a test looks as though it tests what it is supposed to.

Full-time equivalent (FTE) personnel: a hypothetical number of personnel obtained by adding fractional times of personnel actually working. Four half-time persons amount to two people FTE.

Grade-equivalent score: a test score reported as the grade level of an average student who achieves the same performance as the student in question.

Hawthorne Effect: an improvement in performance due merely to the student's having been selected for an evaluation or a special program.

Initial differences: differences in scores between the experimental and control groups on a pretest.

Instructional validity: the extent to which the subject matter of a test is part of students' actual instruction; compare curricular validity.

Instrument: test.

Item: a test question; if multiple choice, the item includes the alternative answers.

Item analysis: in test development, the process of identifying particular items whose removal or revision will improve the test.

Kurtosis: describes a score distribution that is either sharper or squarer than the normal distribution.

Mean: a numerical average.

Median: the middlemost score among all students in the group; generally a better expression of "typical performance" than the mean.

Minimum competency test: a test mandated at the state, county or district level ostensibly to determine whether a high school student is sufficiently competent in basic skills to meet the demands of adult life; also, a test for lower grades that examines basic skills and their application. In some jurisdictions, promotion from grade to grade or receipt of a high school diploma requires certain minimum competency scores.

Mode: the score that the most students achieved.

Multiple-choice test: a test whose questions are accompanied by alternative answers from among which the child chooses one. Distinguished from constructed-response test.

Norm-referenced test: a test whose scores are expressed in terms of the student's standing among a representative sample, as in percentiles, stanines, grade equivalents or corrected scores. Distinguished from criterion-referenced test.

Normal distribution: a particular mathematical distribution of test scores; the "bell-shaped curve."

Norms: tables of test scores achieved by a representative sample of students and organized by percentiles, stanines, etc.; used for converting a raw score to a norm referenced score. (The term "norm" is sometimes also used to mean "average.")

Objective measurement: a form of measurement unaffected by subjective considerations on the part of the examiner; often used synonymously with multiple-choice test.

Objectives: (1) the instructional goals of a curriculum. (2) the instructional goals a test purports to measure. Often objectives for a test are necessarily more restricted than the underlying instructional objectives.

Percentile: the one-hundredth of the group in which a particular student falls, counting from the bottom.

Pilot evaluation: a small-scale evaluation conducted before the main effort in order to try out tests, forms, procedures, etc.

Post-test: a test given after a student's participation in a program, for comparison with a score from an earlier pre-test.

Predictive validity: the extent to which the results from a test agree with those from other measures taken at a later time.

Pre-post gain: the difference in scores between pre-test and post-test, often taken as a measure of learning in a program.

Pre-test: a test given before a student's participation in a program, for comparison with score from a later post-test.

Program evaluation: assessment of an instructional program to judge its instructional effectiveness.

Psychometrician: an expert in constructing tests.

Psychometrics: the science of testing, especially multiple-choice testing.

Qualitative methods: evaluation methods that do not directly yield numerical results; includes interviews, classroom observation, etc.

Random selection: the separation of children into experimental and control groups by choosing the groups at random.

Rating scale: a paper-and-pencil device for obtaining information on opinions. The respondent chooses the point that best represents his opinion between two extremes on the scale.

Raw score: the number of test items a student answers correctly.

Reliability: the extent to which a test gives consistent results. The lower the reliability, the greater the probability of error in the results.

Reliability coefficient: a numerical indication of a test's reliability. Reliability coefficients can range from 0.00 to 1:00; higher is better.

Self-selection: the separation of children into experimental and control groups according to their own or their parents' preferences.

Sensitivity at the extremes: an extra likelihood of error in very high and very low scores.

Significant difference: a difference between two test scores (or two averages of test scores) great enough to be statistically significant under the circumstances.

Skew: describes a score distribution in which scores are clustered near one end and trail off toward the other end.

Split-half analysis: a method of assessing test reliability that compares scores from different batches of questions on the same test.

Standardized test: (1) a test given to students under uniform conditions of time, materials, instructions, etc. (2) a test scored in terms of performance of a representative group of students, or standardization group. In the second sense, the term is almost synonymous with "norm-referenced test."

Stanine: the one-ninth of the group in which a particular student falls, counting from the bottom.

Statistical significance: the unlikelihood that a particular result is due to chance.

Teaching to the test: instruction based so closely on a test as to improve children's performance on that particular test without improving the skills supposedly being tested.

Test anxiety: an anxiety about tests severe enough to impair performance.

Test bias: the extent to which a test is sensitive to children's racial or ethnic backgrounds.

Test/retest: a method of assessing test reliability that compares scores from the same test given on two different occasions.

Test-taking skills: skills, apart from those supposedly being tested, that affect test performance on all tests of the same general type (e.g., multiple choice).

Validity: the extent to which a test actually measures what it claims to measure. See also face validity, content validity, concurrent validity, predictive validity, construct validity, curricular validity, instructional validity.

Validity coefficient: a numerical indication of a test's concurrent or predictive validity, ranging between 0.00 and 1.00. Higher is better.

Acknowledgments

AASA extends its appreciation to Mitchell Lazarus, author of *Evaluating Educational Programs*. Lazarus, who holds a Ph.D. degree in psychology from MIT, is an educational consultant.

We also appreciate the contribution of Robert Bender, Coordinator of Planning for the Planning and Assessment Office of the Fairfax County Public Schools in Virginia, who reviewed the initial manuscript and made substantive comments which have been included in this book. The final chapter was prepared with the assistance of Kristen Amundson, formerly with the National Association of Elementary School Principals, now a communications consultant in the Washington, D.C., area.

Our special thanks to William Spady, Director of AASA's National Center for the Improvement of Learning (NCIL), who offered advice and counsel, and to the hundreds of AASA members who replied to our AASA evaluation survey.

Project direction, editing and final production were under the supervision of AASA Associate Editor, Judy Touchton, and Associate Executive Director, Gary Marx.